3GPP TSG-WG SA2 Meeting #161 *S2-2403335*

Athens, Feb 26th – Mar 1st, 2024

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *CR-Form-v12.2* | | | | | | | | |
| **CHANGE REQUEST** | | | | | | | | |
|  | | | | | | | | |
|  | **23.288** | **CR** | **0996** | **rev** | **4** | **Current version:** | **18.4.0** |  |
|  | | | | | | | | |
| *For* [***HE******LP***](http://www.3gpp.org/3G_Specs/CRs.htm#_blank)*on using this form: comprehensive instructions can be found at* [*http://www.3gpp.org/Change-Requests*](http://www.3gpp.org/Change-Requests)*.* | | | | | | | | |
|  | | | | | | | | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Proposed change affects:*** | UICC apps |  | ME |  | Radio Access Network |  | Core Network | **x** |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | |
| ***Title:*** |  | | | | | | | | | |
|  |  | | | | | | | | | |
| ***Source to WG:*** | Nokia, Nokia Shanghai-Bell, ZTE | | | | | | | | | |
| ***Source to TSG:*** | S2 | | | | | | | | | |
|  |  | | | | | | | | | |
| ***Work item code:*** | eNA\_Ph3 | | | | |  | ***Date:*** | | | 2024-01-12 |
|  |  | | | |  | |  | | |  |
| ***Category:*** | **F** |  | | | | | ***Release:*** | | | Rel-18 |
|  | *Use one of the following categories:* ***F*** *(correction)* ***A*** *(mirror corresponding to a change in an earlier release)* ***B*** *(addition of feature),* ***C*** *(functional modification of feature)* ***D*** *(editorial modification)*  Detailed explanations of the above categories can be found in 3GPP [TR 21.900](http://www.3gpp.org/ftp/Specs/html-info/21900.htm). | | | | | | | | *Use one of the following releases: Rel-8 (Release 8) Rel-9 (Release 9) Rel-10 (Release 10) Rel-11 (Release 11) … Rel-16 (Release 16) Rel-17 (Release 17) Rel-18 (Release 18) Rel-19 (Release 19)* | |
|  |  | | | | | | | | | |
| ***Reason for change:*** | | The term “accuracy” is used to describe the reported accuracy. It is defined in Clause 5c (“Analytics/ML Model Accuracy Monitoring Functional Description”) as follows:  *Analytics/ML Model Accuracy Monitoring is to be achieved by comparing the predictions using the current trained ML model and its corresponding ground truth data i.e. the corresponding true observed events.*  *Analytics/ML Model Accuracy information is to represent general performance measurements for analytics and ML Model respectively, which are composed of the number of correct predictions out of all predictions and the corresponding number of samples.*  *NOTE 2: How MTLF/AnLF determines whether the prediction is correct one is up to implementation.*  Clause 5.C relates to accuracy both of Analytivs and ML models.  However, while the term “accuracy” is used consistenly for analytics, for ML model the term “metrics” appears instead in some places, and with variations of the meaning:   * It can indicate (e.g. in requests) that accuracy information or reporting is requested (the term” ML Model Accuracy Check Flag” is used elsewhere) * it can indicate (e.g. in responses or notifications) the observed (accuracy) value * it can indicate (e.g. in responses or notifications) the (“accuracy”) method (only one in this release) to calculate the observed (accuracy) value   The terminology is a left-over of proposals to support multiple diffrent metrics (With the consumer being allowed to choose the applicable metric in requests) to meassure the correctness of analytivs and models (with “accuracy” based on its current definition beeing one such metric). Those proposals were not agreed, but it is desired to maintain metric for future extensibility | | | | | | | | |
|  | |  | | | | | | | | |
| ***Summary of change:*** | | Consisten usage of terms for models keeping metric for future compatibility:  Usage terminology based on Clause 6.2A.2 with slight updates:  (motivation is to keep overall impacts as small as possible)  *The consumers of the ML model provisioning services (i.e. an NWDAF containing AnLF) may provide*  *- [OPTIONAL] ML Model Monitoring Information:*  *- desired ML Model metric.*  *NOTE 4: In this release, only “ML Model Accuracy” is defined as ML model metric*  *The NWDAF containing MTLF provides to the consumer*  *- [OPTIONAL]ML Model Accuracy Information: indicates the* ***Accuracy*** *of the ML model …, which includes:*  *- the metric value of the ML model.*  *- [OPTIONAL] used ML model metric).* | | | | | | | | |
|  | |  | | | | | | | | |
| ***Consequences if not approved:*** | | Confusing and ambiguous terminology. | | | | | | | | |
|  | |  | | | | | | | | |
| ***Clauses affected:*** | | 6.2A.2, 6.2C.2.2, 6.2C.2.3, 6.2E.3.3, 6.2F.1, 6.2F.2, 6.2F.3, 7.5.2, 7.9.2, 7.9.4, 7.10.4, 7.11.2 | | | | | | | | |
|  | |  | | | | | | | | |
|  | | **Y** | **N** |  | | | |  | | |
| ***Other specs*** | |  | **x** | Other core specifications | | | | TS/TR ... CR ... | | |
| ***affected:*** | |  | **x** | Test specifications | | | | TS/TR ... CR ... | | |
| ***(show related CRs)*** | |  | **x** | O&M Specifications | | | | TS/TR ... CR ... | | |
|  | |  | | | | | | | | |
| ***Other comments:*** | |  | | | | | | | | |
|  | |  | | | | | | | | |
| ***This CR's revision history:*** | |  | | | | | | | | |

1st change

### 6.2A.2 Contents of ML Model Provisioning

The consumers of the ML model provisioning services (i.e. an NWDAF containing AnLF) as described in clause 7.5 and clause 7.6 may provide the input parameters as listed below:

- Information of the analytics for which the requested ML model is to be used, including:

- A list of Analytics ID(s): identifies the analytics for which the ML model is used.

- [OPTIONAL] NF consumer information: identifies the vendor of NWDAF containing AnLF.

NOTE 1: NF consumer information such as Vendor ID is defined in Stage 3.

- [OPTIONAL] Use case context: indicates the context of use of the analytics to select the most relevant ML model ML model.

NOTE 2: The NWDAF containing MTLF can use the parameter "Use case context" to select the most relevant ML model, when several ML models are available for the requested Analytics ID(s). The values of this parameter are not standardized.

- [OPTIONAL] ML Model Interoperability Information. This is vendor-specific information that conveys, e.g., requested model file format, model execution environment, etc. The encoding, format, and value of ML Model Interoperable Information is not specified since it is vendor specific information, and is agreed between vendors, if necessary for sharing purposes.

- [OPTIONAL] ML Model Filter Information: enables the NWDAF containing MTLF to select which ML model for the analytics is requested, e.g. S-NSSAI, Area of Interest. Parameter types in the ML Model Filter Information are the same as parameter types in the Analytics Filter Information which are defined in procedures.

- [OPTIONAL] Target of ML Model Reporting: indicates the object(s) for which ML model is requested, e.g. specific UEs, a group of UE(s) or any UE (i.e. all UEs).

- [OPTIONAL] Requested representative ratio: a minimum percentage of UEs in the group whose data is a non-empty set and can be used in the model training when the Target of ML Model Reporting is a group of UEs.

- ML Model Reporting Information with the following parameters:

- (Only for Nnwdaf\_MLModelProvision\_Subscribe) ML Model Reporting Information Parameters as per Event Reporting Information Parameter defined in Table 4.15.1-1, TS 23.502 [3].

- [OPTIONAL] ML Model Target Period: indicates time interval [start, end] for which ML model for the Analytics is requested. The time interval is expressed with actual start time and actual end time (e.g. via UTC time).

- [OPTIONAL] Inference Input Data information: contains information about various settings that are expected to be used by AnLF during inferences such as:

- the "Input Data" that are expected be used, each of them optionally accompanied by metrics that show the granularity with which this data will be used (i.e., a sampling ratio, the maximum number of input values, and/or a maximum time interval between the samples of this input data).

NOTE 3: This can be a subset of the possible Input Data specified for a certain analytics type.

- the data sources that are expected to be used as a list of NF instance (or NF set) identifiers.

- A Notification Target Address (+ Notification Correlation ID) as defined in clause 4.15.1 of TS 23.502 [3], allowing to correlate notifications received from the NWDAF containing MTLF with this subscription.

- [OPTIONAL] Indication of supporting multiple ML models.

- [OPTIONAL] Accuracy level(s) of Interest.

- [OPTIONAL] Number of ML model(s), indicating the maximum number of ML models that the NWDAF containing MTLF could provide to the NWDAF containing AnLF.

NOTE 4: Multiple ML models Filter Information are composed by Accuracy level(s) of Interest and Number of ML model(s).

- [OPTIONAL] Time when model is needed: indicates the latest time when the consumer expects to receive the ML model(s).

- [OPTIONAL] ML Model Monitoring Information:

- desired ML Model metric.

NOTE 4: In this release, only “ML Model Accuracy” is defined as ML model metric.

- [OPTIONAL] ML model monitoring reporting mode: such as Accuracy reporting interval or pre-determined status. Depending on the reporting mode, the NWDAF containing MTLF reports the model accuracy to NWDAF containing AnLF either periodically or when the ML model accuracy is crossing an ML Model Accuracy threshold, i.e. the accuracy either becomes higher or lower than the ML Model Accuracy threshold.

- [OPTIONAL] ML Model Accuracy Threshold: indicating the accuracy threshold of the ML Model requested by the consumer (as a kind of pre-determined status). It also can be used as an indication that the MTLF is triggered to execute the accuracy monitoring operations for the ML Model provisioned to AnLF.

- [OPTIONAL] DataSetTag and ADRF ID if available: indicates the inference data (including input data, prediction and the ground truth data at the time which the prediction refers to) stored in ADRF which can be used by MTLF to retrain or reprovision of the ML model.

- [OPTIONAL] ML Model Identifier: indicates the Model that the data corresponding to the DataSetTag is related to (in the case of subscription modification).

The NWDAF containing MTLF provides to the consumer of the ML model provisioning service operations as described in clause 7.5 and 7.6, the output information as listed below:

- (Only for Nnwdaf\_MLModelProvision\_Notify) The Notification Correlation Information.

- For each Analytics ID requested by the service consumer, a set of pair (s) of unique ML Model identifier and the following information.

- ML Model Information, which includes:

- the ML model file address (e.g. URL or FQDN); or

- ADRF (Set) ID.

When ADRF (Set) ID is provisioned, a Storage Transaction ID may also be provisioned.

- [OPTIONAL] ML model degradation indicator: indicates whether the provided ML model is degraded.

- [OPTIONAL] Validity period: indicates time period when the provided ML Model Information applies.

- [OPTIONAL] Spatial validity: indicates Area where the provided ML Model Information applies.

NOTE 5: Spatial validity and Validity period are determined by MTLF internal logic and it is a subset of AoI if provided in ML Model Filter Information and of ML Model Target Period, respectively.

- [OPTIONAL] ML model representative ratio: indicating the percentage of UEs in the group whose data is used in the ML model training when the Target of ML Model Reporting is a group of UEs.

- [OPTIONAL] Training Input Data Information: contains information about various settings that have been used by MTLF during training, such as:

- the "Input Data" that have been used, each of them optionally accompanied by metrics that show the data characteristics and granularity with which this data has been used (i.e. a sampling ratio, the maximum number of input values and/or a maximum time interval between the samples of this input data, data range including maximum and minimum values, mean and standard deviation and data distribution when applicable) and the time, i.e. timestamp and duration, when this data was obtained.

- the data sources related to the "Input Data" that were used for ML model training, which have been identified by a list of NF instance (or NF set) identifiers.

NOTE 6: This can be a subset of the possible Input Data specified for a certain analytics type.

NOTE 7: Data source information enables ML Model selection when different models are available for an Analytics ID, or it enables a consumer to avoid selecting a ML model that used data from a specific data source at a particular time or used data characterized by specific data characteristics.

- [OPTIONAL] ML Model Accuracy Information: indicates the accuracy of the ML model if related ML Model Monitoring Information was provided, which includes:

- the metric value of the ML model.

- [OPTIONAL] used ML model metric.

2nd change

#### 6.2C.2.2 General procedure for Federated Learning among Multiple NWDAF Instances



Figure 6.2C.2.2-1: General procedure for Federated Learning among Multiple NWDAF

0. The consumer (NWDAF containing AnLF or NWDAF containing MTLF) sends a subscription request to FL server NWDAF to retrieve an ML model, using Nnwdaf\_MLModelProvision service as defined in clause 7.5 including Analytics ID, desired ML model metric (e.g., ML model Accuracy), Accuracy reporting interval, pre-determined status (ML model Accuracy threshold or Time when the ML model is needed).

NOTE 1: The ML model Accuracy threshold can be used to indicate the target ML Model Accuracy of the training process and the FL server NWDAF may stop the training process when the ML model Accuracy threshold is achieved during the training process.

If the consumer (i.e. the NWDAF containing AnLF or NWDAF containing MTLF) provides the Time when the ML model is needed, the FL Server NWDAF can take this information into account to decide the maximum response time for its FL Client NWDAF(s).

1. FL Server NWDAF selects NWDAF(s) containing MTLF (FL Client NWDAF(s)) as described in clause 6.2C.2.1.

2. FL Server NWDAF sends a Nnwdaf\_MLModelTraining\_Subscribe or Nnwdaf\_MLModelTrainingInfo\_Request to the selected NWDAF(s) containing MTLF (FL Client NWDAF(s)), which participate in the Federated learning to perform the local model training and determine the interim local ML model information based on the input parameter in the request from FL Server NWDAF. The request includes the desired ML model metric and initial ML model and also includes the maximum response time, the FL Client NWDAF has to report the interim local ML model information to the FL Server NWDAF before the maximum response time elapses.

3. [Optional] Each FL Client NWDAF collects its local data by using the current mechanism in clause 6.2 if the Client NWDAF has not local data available already.

4. During Federated Learning training procedure, each FL Client NWDAF further trains the ML model provided by the FL Server NWDAF based on its own data and reports the interim local ML model information to the FL Server NWDAF in Nnwdaf\_MLModelTraining\_Notify or Nnwdaf\_MLModelTrainingInfo\_Request response. The Nnwdaf\_MLModelTraining\_Notify or Nnwdaf\_MLModelTrainingInfo\_Request response may also include the Status report of FL training that includes local ML model metric value (and optionaly the used metric) computed by the FL Client NWDAF and Training Input Data Information (e.g. areas covered by the data set, sampling ratio, maximum/minimum of value of each dimension of data, etc.) in the FL Client NWDAF. The Nnwdaf\_MLModelTraining\_Notify or Nnwdaf\_MLModelTrainingInfo\_Response also includes the global ML Model metric (value (and optionally the used metric) when the ML Model Accuracy Check Flag was included in the Nnwdaf\_MLModelTraining\_Subscribe or Nnwdaf\_MLModelTrainingInfo\_Request (as described in step 7), the global ML Model metric value is calculated by the FL Client NWDAF using the local training data as the testing dataset.

NOTE 2: The parameters in characteristics of local training dataset are up to the implementation.

The local ML model, which is sent from the FL Client NWDAF(s) to the FL Server NWDAF during the FL training process, is the information needed by the FL Server NWDAF to build the aggregated model.

If the FL Client NWDAF is not able to complete the training of the interim local ML model within the maximum response time provided by the FL Server NWDAF, the FL Client NWDAF shall send the Delay Event Notification that include the delay event indication, an optional cause code (e.g. local ML model training failure, more time necessary for local ML model training) and the expected time to complete the training if available to the FL Server NWDAF before the maximum response time elapses.

4a. [Optional]If FL Server NWDAF receives notification/response that the FL Client NWDAF is not able to complete the training within the maximum response time, the FL Server NWDAF may send to the FL Client NWDAF a new maximum response time in Nnwdaf\_MLModelTraining\_Subscribe or Nnwdaf\_MLModelTrainingInfo\_Request, before which the FL Client NWDAF has to report the interim local ML model information to the FL Server NWDAF. Otherwise, the FL Server NWDAF may indicate FL Client NWDAF to skip reporting for this iteration. FL Server NWDAF includes the current iteration round ID in the message to indicate that the request is to modify the training parameters of the current iteration round.

Alternatively, the FL Server NWDAF may inform the FL Client NWDAF to cease the ML model training by sending termination request and to report back the current local ML model updates.

5. The FL Server NWDAF aggregates all the local ML model information retrieved at step 4, to update the global ML model. The FL Server NWDAF may also compute the global ML model metric value, e.g. based on the local ML model metric values(s) or by applying the global model on the validation dataset (if available). The FL Server NWDAF may update the global ML model each time a FL Client NWDAF provides updated local ML model information, or the FL Server NWDAF may decide to wait for local ML model information from all FL Client NWDAFs before updating the global ML model.

If the FL Server NWDAF provides the maximum response time for the FL Client NWDAF(s) to provide the interim local ML model information in step 2, or the new maximum response time in step 4a, the FL Server NWDAF decides either to wait for the FL Client NWDAF(s) which have not yet provided their interim local ML model within the new maximum response time or to aggregate only the retrieved local ML model information instances to update global ML model. The FL Server NWDAF makes this decision, considering the notification/response from the FL Client NWDAF or, if the notification is not received, based on local configuration.

6a. [Optional] Based on the consumer request in step 0, the FL Server NWDAF sends a Nnwdaf\_MLModelProvision\_Notify message to update the ML model metric value to the consumer periodically (e.g. a certain number of training rounds or every 10 min) or dynamically when some pre-determined status is achieved (e.g. the ML Model Accuracy threshold is achieved or training time expires).

6b. [Optional] The consumer decides whether the current model can fulfil the requirement, e.g. global ML model metric value is satisfactory for the consumer and determines to stop or continue the training process. The consumer re-invokes Nnwdaf\_MLModelProvision\_Subscribe service operation as used in step 0 to continue the training process or invokes Nnwdaf\_MLModelProvision\_Unsubscribe service operation to stop the training process.

6c. [Optional] Based on the subscription request sent from the consumer in step 6b, the FL Server NWDAF updates or terminates the current FL training process.

If the FL Server NWDAF received a request in step 6b to stop the Federated Training process, steps 7 and 8 are skipped.

7. If the FL procedure continues, FL Server NWDAF may determine FL Client NWDAF as described in clause 6.2C.2.3 and sends Nnwdaf\_MLModelTraining\_Subscribe or Nnwdaf\_MLModelTrainingInfo\_Request that includes the aggregated ML model information to selected FL Client NWDAF(s) for next round of Federated Training. The request may also include the ML Model Accuracy Check Flag, that indicates the FL Client NWDAF(s) to use the local training data as the testing dataset to calculate the Model Accuracy of the global ML model provided by the FL Server NWDAF.

8. Each FL Client NWDAF updates its own ML model based on the aggregated ML model information distributed by the FL Server NWDAF at step 7.

NOTE 3: The steps 3-8 should be repeated until the training termination condition (e.g. maximum number of iterations, or the result of loss function is lower than a threshold) is reached.

When the Federated Training procedure is complete, the FL Server NWDAF requests the FL client NWDAF(s) to terminate the FL procedure by invoking Nnwdaf\_MLModelTraining\_Unsubscribe service with a cause code that the FL process has finished and optionally with the final aggregated ML model information. Then the FL client NWDAF(s) terminate the local model training and if the final aggregated ML model information is received from the FL server NWDAF, the FL client NWDAF(s) can store it for further use.

9. After the training process is complete, the FL Server NWDAF may send Nnwdaf\_MLModelProvision\_Notify that includes the globally optimal ML model information to the consumer.

3rd change

#### 6.2C.2.3 Procedures for Maintaining Federated Learning Processes

This clause specifies how to maintain a Federation Learning process in FL execution phase, including FL Server NWDAF triggers reselection, addition, or removal of FL Client NWDAF(s), discovers new FL Client NWDAF(s) via NRF and FL Client NWDAF(s) joins or leaves Federated Learning process dynamically.

In Federated Learning execution phase, FL Server NWDAF monitors the status changes of FL Client NWDAF(s) and may reselects FL Client NWDAF(s) based on the updated status, availability and/or capability, etc.

NOTE 1: FL Server NWDAF checks if there is a need to carry on the FL execution phase and then reselects FL members for the next iteration if needed.



Figure 6.2C.2.3-1: Procedure of FL Server NWDAF reselects FL Client NWDAF(s), FL Client NWDAF(s) Join or Leave Federated Learning Process Dynamically in Federated Learning execution phase

The procedure for FL Server NWDAF reselecting FL Client NWDAF(s), FL Client NWDAF(s) joining or leaving Federated Learning process dynamically is as follows:

1a. FL Server NWDAF may get the updated status of current FL Client NWDAF(s) via NRF by using Nnrf\_NFManagement service (as in clause 5.2.7.2 of TS 23.502 [3]) in the Federated Learning execution phase.

FL Server NWDAF may subscribe to NRF for notifications of status changes of the current NWDAF(s) (FL Client NWDAFs 1…N) by invoking an Nnrf\_NFManagement\_NFStatusSubscribe service operation. NRF notifies the FL Server NWDAF the status changes of the current FL Client NWDAF(s) by invoking Nnrf\_NFManagement\_NFStatusNotify service operation(s).

The status of a current FL Client NWDAF could be availability changes, capability changes (e.g. it will not support FL anymore, etc.).

1b. The current FL Client NWDAF(s) may inform FL Server NWDAF that it is leaving the Federated Learning process by invoking Nnwdaf\_MLModelTraining\_Notify service operation with Termination Request and cause code (reason for leaving, e.g. high NF load, time availability changes).

1c. FL Server NWDAF may get the information of the new FL Client NWDAF(s) dynamically via NRF by subscribing to the event that a new FL Client NWDAF registers (Nnrf\_NFManagement\_NFStatusSubscribe service as in clause 5.2.7.2 of TS 23.502 [3]).

1d. NWDAF may subscribe for NF load analytics of the FL Client NWDAF(s).

1e. FL Client NWDAF(s) may send Status report of FL training and Global ML Model Accuracy Information by invoking Nnwdaf\_MLModelTraining\_Notify service.

2. FL Server NWDAF checks FL Client NWDAF(s) status based on the received information and may determine whether reselection of FL Client NWDAF(s) for the next round(s) of Federated Learning is needed based on the received information from step 1.

NOTE 2: Several examples of the factors that the FL Server NWDAF can consider to reselect the FL Client NWDAF(s) are updated status of FL Client NWDAF reported by NRF is different than the criteria were initially used for selecting the client; characteristics of local training dataset is different than global validation dataset owned by FL Server NWDAF and/or the metric value of the global model calculated using the local training dataset is much different from that calculated by other FL Client NWDAFs; the metric value of the global model calculated using the local training dataset is lower than the metric value calculated using the global validation dataset owned by FL Server NWDAF; the metric value of the global model calculated using the local training dataset is lower than ML Model metric value received in Nnwdaf\_MLModelMonitor\_Notify when FL Server NWDAF using AnLF-assisted MTLF ML Models Accuracy Monitoring; the load of the FL Client NWDAF (from the NF load Analytics or from the FL Client NWDAF directly) is high; the FL Server NWDAF receives leave request from the FL Client NWDAF; the FL Client NWDAF did not report the status of FL Training within the maximum response time.

3. [If re-selection is needed as judged in step 2] If step 1c is not performed, FL Server NWDAF may discover new candidate FL Client NWDAF(s) via NRF by using Nnrf\_NFDiscovery services as in clause 5.2.7.3 of TS 23.502 [3]. FL Server NWDAF reselects FL Client NWDAF(s) from the current FL Client NWDAF(s) and the new candidate FL Client NWDAF(s) (found in steps 1c or 3). For the new candidate FL Client NWDAF(s), the interaction between FL Server NWDAF and FL Client NWDAF(s) is same as the selection procedure described in clause 6.2C.2.1. The adding / deleting FL Client NWDAF(s) may happen at the end of each iteration.

4. FL Server NWDAF sends termination request by invoking Nnwdaf\_MLModelTraining\_Unsubscribe service operation or Nnwdaf\_MLModelTrainingInfo\_Request service operation with Correlation Termination Flag to the FL Client NWDAF(s), optionally indicating the reason, e.g. FL Client NWDAF is unselected by the FL Server NWDAF for the FL process, or the FL process is suspended, etc. And FL server may also send the updated global ML model information to the unselected FL client NWDAF. FL Client NWDAF(s) terminates operations for the Federated Learning process if receive termination request from the FL Server NWDAF and may perform further action to be qualified in participation of FL training in the next cycles.

NOTE 3: In the case of high load, the FL Client NWDAF can e.g. decline new service request. In the case of low Accuracy, the FL Client NWDAF can gather new local training data.

4th change

#### 6.2E.3.3 Procedures for monitoring the analytics accuracy of an ML model

An NWDAF containing MTLF, due to the registration of monitoring of the analytics accuracy of an ML model received from NWDAF containing AnLF and local policies, subscribes to the NWDAF containing AnLF for receiving notifications of either the accuracy of the ML Model, or Analytics feedback information of the ML model. The NWDAF containing MTLF may get the Subscription endpoint address of the NWDAF containing AnLF from the information received in a previous registration or through a service discovery procedure at the NRF.

Figure 6.2E.3.3-1 illustrates the procedure either for monitoring the analytics accuracy of an ML model or for delivery of Analytics feedback information of an ML model. Nnwdaf\_MLModelMonitor\_Subscribe and Nnwdaf\_MLModelMonitor\_Notify service operations are used for the purposes. A service consumer, i.e. an NWDAF containing MTLF, subscribes at a service producer, i.e. an NWDAF containing AnLF, to be notified when either the analytics accuracy of the previously provisioned ML model is not sufficient, or Analytics feedback information is retrieved from analytics consumer NF.



Figure 6.2E.3.3-1: Procedure for monitoring the analytics accuracy of an ML model

0. Upon the reception of an Nnwdaf\_MLModelMonitor\_Register request and based on local policy, the NWDAF containing MTLF determines to subscribe to the analytics accuracy monitoring for the ML model as defined in clause 5C.1.

1. The NWDAF containing MTLF sends an Nnwdaf\_MLModelMonitor\_Subscribe request (Analytics ID(s), unique identifier(s) of the ML model(s) to be monitored, desired accuracy metrics to be monitored, optionally Reporting Threshold(s) or Reporting Period) to an NWDAF containing AnLF subscription endpoint.

When the NWDAF containing MTLF determines during the registration process described in clause 6.2E.3.2 that a subscription request for ML model accuracy monitoring to an NWDAF containing AnLF is related to a previous subscription for ML model accuracy information to a different NWDAF containing AnLF (due to changes in the provider of the ML accuracy monitoring for a given ML model, as an effect of analytics transfer among NWDAFs containing AnLF), the NWDAF containing MTLF may use as base for the new subscription request at the new NWDAF containing AnLF the parameters associated with the original subscription identification for the ML model accuracy information that was received in the registration request of the new NWDAF containing AnLF, as described in steps 1-2 of clause 6.2E.3.2.

2. The NWDAF containing AnLF sends a response to the NWDAF containing MTLF.

3. The analytics consumer NF may send Analytics feedback information to the NWDAF containing AnLF as described in clause 6.1.1.

4. When step 1 is triggered, the NWDAF containing AnLF may start monitoring the analytics accuracy of the ML model(s), if it not started yet.

NOTE 1: The NWDAF containing AnLF can monitor the analytics accuracy in many ways: e.g. comparing predictions of ML model and its corresponding ground truth data, comparing changes in internal configuration for the analytics ID generation, previous existent records of analytics accuracy information etc.

5. The NWDAF containing AnLF determines whether the analytics accuracy of the ML model is insufficient, i.e. deviation of the output analytics using the trained ML model from ground truth data (which are collected from Data Producer NF corresponding to analytic ID requested at the time which the prediction refers to) is greater than the Reporting Threshold(s) (which are locally configured or received in the Subscribe request), or the Reporting Period indicated in the Subscribe request is reached.

6. Either the Analytics feedback information is retrieved at step 3 or the NWDAF containing AnLF detects the analytics accuracy of ML model is insufficient at step 5, the NWDAF containing AnLF sends an Nnwdaf\_MLModelMonitor\_Notify request to the notification endpoint (e.g. the NWDAF containing MTLF). The Notify request includes either Analytics feedback information, or the monitored ML model accuracy information of the ML model (e.g. a Deviation value which indicates the deviation of the predictions generated using the ML model(s) from the ground truth data and the network data when the deviation occurs (which can be used by the NWDAF containing MTLF for possible ML model retraining) and the number of inferences that were performed during the time interval between Nnwdaf\_MLModelMonitor\_Register request and the Notify request or between the time of last Notification message and the time of the current Notification message and optionally an indication that the analytics accuracy of the ML model does not meet the requirement of accuracy for the ML model.

7. The NWDAF containing MTLF sends a response.

8. The NWDAF containing MTLF determines whether the ML model is degraded or not based on the notification at step 6. If the notification contains Analytics feedback information, the NWDAF containing MTLF may determine ML model degradation based on the procedures as described in clause 6.2E.2. Otherwise when the NWDAF containing MTLF has received the multiple analytics accuracy information, from one or more NWDAFs containing AnLF, it may consider that the ML model is degraded/to be updated (i.e. enough number analytics accuracy information received from one or more NWDAFs containing AnLF, indicating insufficient analytics accuracy).

NOTE 2: The actual mechanism for the NWDAF containing MTLF for determining the degradation of the ML model degradation is an internal procedure of the NWDAF containing MTLF, e.g. the NWDAF containing MTLF calculate a global accuracy based on the analytics accuracy information and the number of inferences received from multiple NWDAFs containing AnLF.

9. When an ML model is considered degraded / to be updated at step 8, the NWDAF containing MTLF re-trains the existing ML model or selects a new ML model. If the network data was not included in the Nnwdaf\_MLModelMonitor\_Notify request of step 5, the NWDAF containing MTLF may request data from the NWDAF containing AnLF, ADRF and/or other 5GS entities as specified in clause 6.2 and use the collected data for ML model retraining. The NWDAF containing MTLF notifies the NWDAF(s) containing AnLF with the updated trained ML Model Information by invoking Nnwdaf\_MLModelProvision\_Notify service operation, as described in clause 6.2A.

5th change

### 6.2F.1 ML Model Training Subscribe/Unsubscribe

The procedure in Figure 6.2F.1-1 is used by an NWDAF service consumer, i.e. an NWDAF containing MTLF to subscribe to another NWDAF, i.e. an NWDAF containing MTLF, for a trained ML model based on the ML model file or ML Model information as described in clause 6.2F.2 provided by the NWDAF service consumer. The service may be used by an NWDAF containing MTLF to enable e.g. Federated Learning or to update ML model. The service is also used by an NWDAF service consumer to request an NWDAF containing MTLF to prepare training ML model or modify existing ML Model training subscription.



Figure 6.2F.1-1: Procedure for ML Model Training subscribe/unsubscribe

1. The NWDAF service consumer may subscribe or unsubscribe for training an ML model by invoking the Nnwdaf\_MLModelTraining\_Subscribe/ Nnwdaf\_MLModelTraining\_Unsubscribe service operation. The parameters that can be provided by the NWDAF service consumer are listed in clause 6.2F.2.

In order to enable Federated Learning, NWDAF Service consumer act as FL Server NWDAF can subscribe to multiple NWDAFs containing MTLF act as FL Client NWDAFs, which are selected by the FL Server NWDAF.

The FL server NWDAF may use the request to check if an NWDAF can meet the ML model training requirement (e.g. ML Model Interoperability information, Analytics ID, Serving Area and/or availability of data and time). In such case, the FL server NWDAF includes an ML Preparation Flag. When the ML Preparation Flag presents in the request, the service provider NWDAF only checks if it can meet the ML model training requirement (e.g. ML Model Interoperability information, Analytics ID, Serving Area and/or availability of data and time) and / or can successfully download the model if the model information is provided.

The FL server NWDAF may use the request to get the Model Accuracy information of the global ML Model calculated by the FL Client NWDAFs. In such cases, the service consumer NWDAF includes a Model Accuracy Check Flag. When the Model Accuracy Check Flag is present in the request, the service provider NWDAF uses the local training data as the testing dataset to calculate the Model Accuracy information of the ML model provided by the service consumer NWDAF.

When NWDAF service consumer determine to further update the ML model, NWDAF service consumer modifies the subscription by invoking Nnwdaf\_MLModelTraining\_Subscribe service operation including Subscription Correlation ID with ML Model Information (as defined in clause 6.2A.2).

2. The NWDAF containing MTLF trains ML model provided at step 1 by collecting new data or re-use the data that it owns. If the ML model file is not provided in step 1, the NWDAF containing MTLF shall first get the ML model using the information indicated at step 1.

3. When the NWDAF containing MTLF completes ML model training, the NWDAF containing MTLF notifies the NWDAF service consumer with ML Model Information (as defined in clause 6.2A.2) of updated ML Model) by invoking the Nnwdaf\_MLModelTraining\_Notify service operation. The parameters that can be provided by the NWDAF containing MTLF as service provider is specified in clause 6.2F.2.

If the NWDAF containing MTLF determines to terminate the ML model training, i.e. NWDAF containing MTLF will not provide further notifications related to this request, then the NWDAF containing MTLF may notify the NWDAF Service consumer a Terminate Request indication with cause code (e.g. NWDAF overload, not available for the FL process anymore, etc.) by invoking the Nnwdaf\_MLModelTraining\_Notify service operation.

In order to enable Federated Learning, NWDAF containing MTLF acting as FL Client NWDAF can notify NWDAF Service consumer acting as FL Server NWDAF the local ML model information and status report of FL training including Accuracy information of local model and Training Input Data Information (e.g. areas covered by the data set, sampling ratio, maximum/minimum of value of each dimension, etc.).

If the Model Accuracy Check Flag is present in the Nnwdaf\_MLModelTraining\_Subscribe, the service provider NWDAF acting as FL Client NWDAF may notify the NWDAF Service consumer acting as FL Server NWDAF the Model Accuracy information of the global ML Model.

6th change

### 6.2F.2 Contents of ML Model Training

The consumers of the ML model training services (i.e. an NWDAF containing MTLF) may provide the input parameters in Nnwdaf\_MLModelTraining\_Subscribe or Nnwdaf\_MLModelTrainingInfo\_Request service operations as listed below:

- Analytics ID: identifies the analytics for which the ML model is requested to be trained.

- ML Model Interoperability Information as defined in clause 6.2A.2.

- (Only for Nnwdaf\_MLModelTraining\_Subscribe) A Notification Target Address (+ Notification Correlation ID) as defined in TS 23.502 [3] clause 4.15.1, allowing to correlate notifications received from the NWDAF containing MTLF with the subscription.

- [OPTIONAL] ML Model Information (as defined in clause 6.2A.2).

- [OPTIONAL] ML Model file.

NOTE 1: It is up to NWDAF implementation to determine whether to include ML Model file in input parameters considering ML Model file size, etc.

- [OPTIONAL] ML Model ID: identifies the provided ML model.

- [OPTIONAL] ML Preparation Flag: identifies whether the request is for preparing Federated Learning or executing Federated Learning.

- [OPTIONAL] ML Model Accuracy Check Flag: identifies that the request is for using the local training data as the testing dataset to calculate the Model Accuracy of the global ML model provided by the NWDAF service consumer acting as the FL Server NWDAF.

- [OPTIONAL] ML Correlation ID: identifies the Federated Learning procedure for training the ML model. This parameter is included when the service is used for Federated Learning.

- [OPTIONAL] Available data requirement. This is for informing the requirement on available data for the ML model training. e.g. FL Server NWDAF sends the requirement in preparation request to a FL Client NWDAF for selecting the FL Client NWDAF which can meet the available data requirement. The following available data requirements can be included:

- Event ID list to be collected for local model training.

- Dataset statistical properties as defined in clause 6.1.3.

- Time window of the data samples.

- Minimum number of data samples.

- [OPTIONAL] Availability time requirement. This is for informing the requirement on availability time for the ML model training, e.g. FL Server NWDAF sends the requirement in preparation request to FL Client NWDAF for selecting the FL Client NWDAF which is available in the required time for training ML model.

- [OPTIONAL] Training Filter Information: enables to select which data for the ML model training is requested, e.g. S-NSSAI, Area of Interest. Parameter types in the Training Filter Information are the same as or subset of parameter types in the ML Model Filter Information which are defined in procedure 6.2A.1.

- [OPTIONAL] Target of Training Reporting: indicates the object(s) for which data for ML model training is requested, i.e. a group of UEs or any UE (i.e. all UEs).

- [OPTIONAL] Use case context: indicates the context of use of ML model.

- [OPTIONAL] Training Reporting Information with the following parameters:

- Maximum response time: indicates maximum time for waiting notifications (i.e. model training results).

- [OPTIONAL] Iteration round ID: indicates the iteration round number of current ML model training.

- [OPTIONAL] Expiry time.

The NWDAF containing MTLF provides to the consumer of the ML model training service operations as described in clause 7.10, the output information in notification as listed below:

- The Notification Correlation Information.

- [OPTIONAL] ML Model Information (as defined in clause 6.2A.2).

- [OPTIONAL] ML Model ID: identifies the provisioned ML model.

- [OPTIONAL] Model Accuracy information: The model metric value of the global ML model and optionally the used metric, which is calculate by the FL Client NWDAF using the local training data as the testing dataset.

[OPTIONAL] Status report of FL training: Accuracy information of local model and Training Input Data Information (e.g. areas covered by the data set, sampling ratio, maximum/minimum of value of each dimension, etc.), which are generated by the FL Client NWDAF during FL procedure.

NOTE 2: The parameters in Training Input Data Information are up to the implementation.

- [OPTIONAL] ML Correlation ID. This parameter may be included when the service is used for Federated Learning.

- [OPTIONAL] Iteration round ID: indicates the iteration round number of ML model training indicated by the FL Server NWDAF.

- [OPTIONAL] Delay Event Notification with the following parameters:

- delay event indication: this parameter indicates that FL Client NWDAF is not able to complete the training of the interim local ML model within the maximum response time provided by the FL Server NWDAF.

- [OPTIONAL] cause code (e.g. local ML model training failure, more time necessary for local ML model training, etc.).

- [OPTIONAL] Expected time to complete the training: Indicates to the FL Server NWDAF that expected remaining training time and may be provided with Delay Event Notification.

7th change

### 6.2F.3 ML Model Training Information Request

The procedure in Figure 6.2F.3-1 is used by an NWDAF service consumer, i.e., an NWDAF containing MTLF to request another NWDAF, i.e., an NWDAF containing MTLF, for the information about ML model training based on the ML model file or ML Model information as described in clause 6.2F.2 provided by the NWDAF service consumer. The service may be used by an NWDAF containing MTLF to enable e.g. Federated Learning.



Figure 6.2F.3-1: Procedure for ML Model Training Information Request

1. The NWDAF service consumer may request the NWDAF containing MTLF to get the information about the ML model training based on the ML model file or ML Model information as described in clause 6.2F.2 provided by the service consumer by invoking the Nnwdaf\_MLModelTrainingInfo\_Request service operation. The parameters that can be provided by the NWDAF service consumer are listed in clause 6.2F.2.

In order to enable Federated Learning, NWDAF Service consumer acting as FL Server NWDAF requests to get ML Model Training Information from an NWDAF containing MTLF acting as FL Client NWDAF, which is selected by the FL Server NWDAF. The details are specified in clause 6.2C.

The NWDAF service consumer may use the request to check if an NWDAF can meet the ML model training requirements (e.g. ML Model Interoperability information, Analytics ID, Service Area/DNAI and/or availability of data and time). In such cases, the NWDAF service consumer includes an ML Preparation Flag.

The NWDAF service consumer may use the request to get the Model Accuracy of the ML Model provided by the service consumer using local training data in the NWDAF containing MTLF as the testing dataset. In such cases, the service consumer NWDAF includes a Model Accuracy Check Flag.

2. When the ML Preparation Flag is present in the request, the NWDAF containing MTLF only checks whether it can meet the ML model training requirement and/or can successfully download the model if the model information is provided. Based on the check result, the NWDAF containing MTLF gets a successful return code or failure cause code (e.g. NWDAF does not meet the ML training requirements) as the information about the ML model training.

When the Model Accuracy Check Flag is present in the request, the NWDAF containing MTLF uses the local training data as the testing dataset to calculate the Model Accuracy information of the ML model provided by the service consumer. The NWDAF containing MTLF includes the Model Accuracy information into the information about the ML model training.

When the NWDAF containing MTLF is ongoing ML model training based on the ML model file or ML Model information as described in clause 6.2F.2 provided by the NWDAF service consumer, the NWDAF containing MTLF gets a failure cause code (e.g. ML training is not complete) as the information about the ML model training.

When the NWDAF containing MTLF completes ML model training based on the ML model file or ML Model information as described in clause 6.2F.2 provided by the NWDAF service consumer, the NWDAF containing MTLF gets a successful return code and the ML Model Information of the trained ML model as the information about the ML model training.

3. The NWDAF containing MTLF replies to the NWDAF service consumer with the information about the ML model training by invoking the Nnwdaf\_MLModelTrainingInfo\_Request response service operation.

8th change

### 7.5.2 Nnwdaf\_MLModelProvision\_Subscribe service operation

**Service operation name:** Nnwdaf\_MLModelProvision\_Subscribe.

**Description:** Subscribes to NWDAF ML model provision with specific parameters.

**Inputs, Required:** (set of) Analytics ID(s) defined in Table 7.1-2, Notification Target Address (+ Notification Correlation ID).

**Inputs, Optional:** Subscription Correlation ID (in the case of modification of the ML model subscription), ML Model Filter Information to indicate the conditions for which ML model for the analytics is requested and Target of ML Model Reporting to indicate the object(s) for which ML model is requested (e.g. specific UEs, a group of UE(s) or any UE (i.e. all UEs)), NF consumer information, Requested representative ratio, ML Model Reporting Information (including e.g. ML Model Target Period), Expiry time, Use case context, Inference Input Data information, indication of support for multiple ML models, multiple ML models Filter Information to indicate the conditions for which multiple ML models are requested, ML Model Interoperability Information, Time when model is needed, ML Model Monitoring Information (including e.g. desired ML Model metric, ML model monitoring reporting mode, ML Model Accuracy Threshold, DataSetTag and ADRF ID, ML Model Identifier).

**Outputs Required:** When the subscription is accepted: Subscription Correlation ID (required for management of this subscription), Expiry time (required if the subscription can be expired based on the operator's policy).

**Outputs, Optional:** None.

9th change

### 7.9.2 Nnwdaf\_MLModelMonitor\_Subscribe service operation

**Service operation name:** Nnwdaf\_MLModelMonitor\_Subscribe

**Description:** Subscribes to NWDAF for ML model accuracy (i.e. Analytics accuracy for an ML model as described in clause 6.2E.3.3) information and Analytics feedback information for the analytics generated by the NWDAF with specific parameters.

**Inputs, Required:** (set of) Unique ML Model identifier(s), Notification Target Address (+ Notification Correlation ID).

**Inputs, Optional:** Subscription Correlation ID (in the case of modification of the ML model monitor subscription), desired Accuracy metrics to indicate the metrics to calculate the accuracy information, reporting period to indicate the reporting periodicity in which the monitored ML Model accuracy information shall be reported, Accuracy reporting threshold to indicate the reporting condition above which the accuracy information shall be reported.

**Outputs Required:** When the subscription is accepted: Subscription Correlation ID (required for management of this subscription), Expiry time (required if the subscription can be expired based on the operator's policy).

**Outputs, Optional:** None.

10th change

### 7.9.4 Nnwdaf\_MLModelMonitor\_Notify service operation

**Service operation name:** Nnwdaf\_MLModelMonitor\_Notify.

**Description:** NWDAF notifies the monitored ML model accuracy (i.e. Analytics accuracy for an ML model as described in clause 6.2E.3.3) information and Analytics feedback information for the analytics generated by the NWDAF to the consumer instance which has subscribed to the specific NWDAF service.

**Inputs, Required:** Notification Correlation Information, at least one of the following:

- the monitored ML model accuracy information which includes:

- Unique ML model identifier;

- Monitoring interval: time interval during which the ML model accuracy monitoring was conducted;

- Monitored Analytics metrics value of the ML Model and a deviation value which indicates the deviation of the predictions generated using the ML model(s) from the ground truth data;

- Number of inferences that were performed during the monitoring interval;

- used Accuracy metrics (as requested in Subscribe service operation).

- Analytics feedback information: indicates that the consumer NF of the analytics generated by the provisioned ML model has taken an action(s) influenced by the analytics and includes the following parameter(s):

- Corresponding Analytics ID(s) which has been used for taking an action(s);

- Corresponding ML Model identifier(s) which has been used for generating Analytics;

- Indication whether the action will affect on ground truth data (if available);

- Time stamp(s) when the action(s) are taken.

**Inputs, Optional:**

- Input data used for inferencing indicated by DataSetTag with ADRF ID when the prediction generated from the ML Model is not correct (which can be used by the NWDAF containing MTLF for possible ML model retraining);

NOTE: How MTLF/AnLF determines whether the prediction is correct one is up to implementation.

- An indication that the analytics metrics value of the ML model does not meet the requirement of accuracy for the ML model.

**Outputs, Required:** Operation execution result indication.

**Outputs, Optional:** None.

11th change

### 7.10.4 Nnwdaf\_MLModelTraining\_Notify service operation

**Service operation name:** Nnwdaf\_MLModelTraining\_Notify

**Description:** NWDAF notifies the consumer instance of the trained ML model that has subscribed to the specific NWDAF service. The NWDAF can also use this service to indicate to consumer it will terminate the ML model training.

**Inputs, Required:**

- Notification Correlation Information: this parameter indicates the Notification Correlation ID that has been assigned by the consumer during ML model training.

**Inputs, Optional:**

- Set of the tuple (Analytics ID, ML model Information as defined in clause 6.2F.2;

- ML Correlation ID, when for Federated Learning;

- Corresponding Use case context;

- Termination Request: this parameter indicates that NWDAF requests to terminate the ML model training, i.e. NWDAF will not provide further notifications related to this request, with cause code (e.g. NWDAF overload, not available for the FL process anymore, etc.);

- ML Model ID: this parameter identifies the provisioned ML model;

- Global ML Model Accuracy information: The model accuracy of the global ML model, which is calculate by the FL Client NWDAF using the local training data as the testing dataset;

- Status report of FL training: local ML Model accuracy information and Training Input Data Information (e.g. areas covered by the data set, sampling ratio, maximum/minimum of value of each dimension, etc.), which are generated by the FL Client NWDAF during FL procedure;

- Delay Event Notification: as defined in clause 6.2F.2;

- Iteration round ID.

NOTE: The detail reasons in the cause code are up to stage 3.

**Outputs, Required:** Operation execution result indication.

**Outputs, Optional:** None.

12th change

### 7.11.2 Nnwdaf\_MLModelTrainingInfo\_Request service operation

**Service operation name:** Nnwdaf\_MLModelTrainingInfo\_Request

**Description:** Request information about NWDAF ML model training with specific parameters.

**Inputs, Required:**

- Analytics ID as defined in Table 7.1-2.

- ML Model Interoperability information.

**Inputs, Optional:**

- ML Model ID: identifies the provided ML model.

- ML Model Information (as defined in clause 6.2A.2).

- ML Model file.

- ML Training Information (i.e. data availability requirement, time availability requirement).

- Training Reporting Information as defined in clause 6.2F.2.

- ML Preparation Flag.

- ML Model Accuracy Check Flag.

- ML Correlation ID.

- Termination Request, when terminating the Federated Learning identified by the ML Correlation ID and optionally indicating the reason, e.g. FL Client NWDAF is unselected by the FL Server NWDAF for the FL process, or the FL process is suspended, etc.

- Training Filter Information.

- Target of Training Reporting.

- Use case context.

**Outputs Required:** When the request is accepted: Operation execution result indication. When the request is not accepted, an error response with cause code (e.g. NWDAF does not meet the ML training requirements, ML training is not complete, NWDAF overload, not available for the FL process anymore, etc.).

NOTE: The detail reasons in the cause code are up to stage 3.

**Outputs, Optional:**

- ML Model ID.

- Set of the tuple (Analytics ID, ML model Information (as defined in clause 6.2A.2)).

- ML Correlation ID, when for Federated Learning.

- Corresponding Use case context.

- Global ML Model Accuracy information: The model accuracy of the global ML model, which is calculate by the FL Client NWDAF using the local training data as the testing dataset.

- Status report of FL training: local ML model accuracy information and Training Input Data Information (e.g. areas covered by the data set, sampling ratio, maximum/minimum of value of each dimension of data, etc.), which are generated by the FL Client NWDAF during FL procedure.

- Delay Event Notification as defined in clause 6.2F.2;

End of changes