



Meteo Italian Supercomputing poRtAL


Deliverable

D5.7 High Availability Requirement

Deliverable Lead:	Arpa Emilia Romagna
Deliverable due date	30/sept/2019
Version	FINAL
Status	V1

Document Control Page

Title	D5.7 High Availability Requirement
Creator	Davide Cesari
Publisher	Mistral Consortium
Contributors	
Type	<<Report.>>
Language	en-GB
Rights	copyright "Mistral Consortium"
Audience	<input checked="" type="checkbox"/> public <input type="checkbox"/> restricted
Requested deadline	



Executive Summary	4
Introduction	5
Problems related to redundancy of data	5
Possible mitigating solutions	6
Criteria for tolerable delays	7

Executive Summary

D5.7 is the description for Subtask 5.3.1: High Availability Requirement Analysis for forecast data.

With this task, we examine the different access patterns to data in high availability and define the requirements for the access to these data.

High Availability Requirement

High Availability Requirement Analysis for forecast data will examine the different access patterns to data in high availability and define the requirements for the access to these data, e.g. in terms of tolerable delays and tolerable increased complexity of access; it will define the strategies for handling the different sources of data and managing the cases of failure of one of them.

Introduction

Some forecast datasets hosted in the Mistral Data Portal may be produced simultaneously by more than one equivalent real-time procedures running on independent High-Performance Computing (HPC) systems. The independent systems may be colocated in the same data center, as it is the case now in Cineca for the COSMO-LAMI Numerical Weather Prediction (NWP) procedures, or they may be distributed in different locations and possibly managed by different organisations.

This redundancy is designed for guaranteeing the high availability (HA) of data, but it involves additional problems in the acquisition of the data by the Data Portal archive.

Problems related to redundancy of data

Due to the nature of NWP models and HPC systems, in case of redundant generation of forecasts, it is not advisable to mix data for a single model run from different computing systems, because, even when maximum care is taken in order to implement the procedures identically in the different systems, using the same software version and configuration, same compilers and compiler options, etc. slight differences in the results are always possible, e.g. due to different floating-point models and non-reproducibility of algorithms especially regarding parallel processing. These differences, although frequently lying well below the range of model error, can be strongly amplified when calculating time-derived quantities, e.g. accumulated precipitation, where unphysical negative values or other kinds of artefacts may appear.

The first consequence of this fact is that, if a forecast dataset has more than one different redundant sources, the data relative to a single model run, for that dataset, can be published only when the fastest redundant source of data has sent all the data for that run.

On the archiving side, it has to be taken into account that, before a real-time NWP model dataset is ready for download by an end user, a big amount of data has to be transferred (possibly from a remote system), stored, indexed and possibly undergo a preliminary postprocessing for generating some common output that is required by many users but is not present in the raw model output. These operations of transfer, archiving and processing require a significant amount of CPU and I/O resources. If particular care is not taken, the time for performing these operations may become comparable to the time required for generating the forecast itself on an HPC system, with the consequence of significant investments in HPC resources.

We are thus in front of two competing factors: on the one hand, the duplication of data sources for improving high availability of forecasts has the consequence that, until at least one model run has not terminated, it is not clear which data source will be the final data source. On the other hand, the transfer, processing and archiving of data has to start while the model is still running in order not to add delays in the availability of the complete datasets.

Possible mitigating solutions

For the reasons indicated above, all the data sources should be transferred to the Data Portal and processed as soon as the model has started producing data, without waiting for the forecast to be complete.

Furthermore, since the archiving and indexing of the data in the archive involves merging huge files containing new data with existing files containing data archived after the previous model runs, the archiving process is not instantaneous and not easily reversible. For speeding up this step of merging the archives, a couple of solutions are proposed:

1. at a system level, for example, using zfs or btrfs filesystems, create several copies or clones of the involved dataset equal to the number of HA data sources and populate each of them with the corresponding source as soon as new data become available; when the first data source has been completely imported, the primary dataset can be (almost) atomically replaced by the clone and the other clone(s) can be destroyed.
2. at a software level, e.g. in the arkimet archiving software, implement an incremental archiving feature which speeds up the merge and combined indexing of two files belonging to the same segment of a dataset; the HA sources are imported separately, each one in an own temporary file; when the first data source has been completely imported it is merged with the main dataset using the incremental archiving feature.

Alternatively, if the system is performant enough, it could be acceptable to transfer and process the different HA data sources in parallel while the model is running and start archiving and indexing the first one that becomes available just after it is complete.

Criteria for tolerable delays

The following criterion is defined for considering acceptable an archiving method: the delay between the end of the forecast on the HPC system where the forecast is produced and the time at which data are available to the users in the Data Portal must not exceed the 5% of the time required for running the forecast in the HPC system.

However, in no cases, the measures taken for speeding up the availability of data should decrease the reliability of all the processing chain.