ON THE POTENTIAL OF BIG DATA CAPABILITIES FOR THE VALIDATION OF A WEATHER FORECASTING SYSTEM

Giuseppe Iannitto, Università degli Studi di Roma Tor Vergata, EO Lab

Outline

- Background – Instant Weather app
- The validation process
- Big data issues and solutions
- The proposed solution
- Conclusions

Background

Introduction

- A key component of the EO projects is the validation of the EO data products through a Ground Truth Validation.
- In the validation process data can be collected from various ground-based sources and sensors (in situ measurements, instruments, crowd-sourcing, open source platform), then quality-controlled, and finally compared with the satellite products in order to get validated retrievals.
- The objective of this work is to develop a system that uses big data capabilities and tools for validation purposes, in particular for the assessment of a new weather nowcasting system, based on a predictive model exploiting Meteosat Second Generation (MSG) imagery.

The Instant Weather app

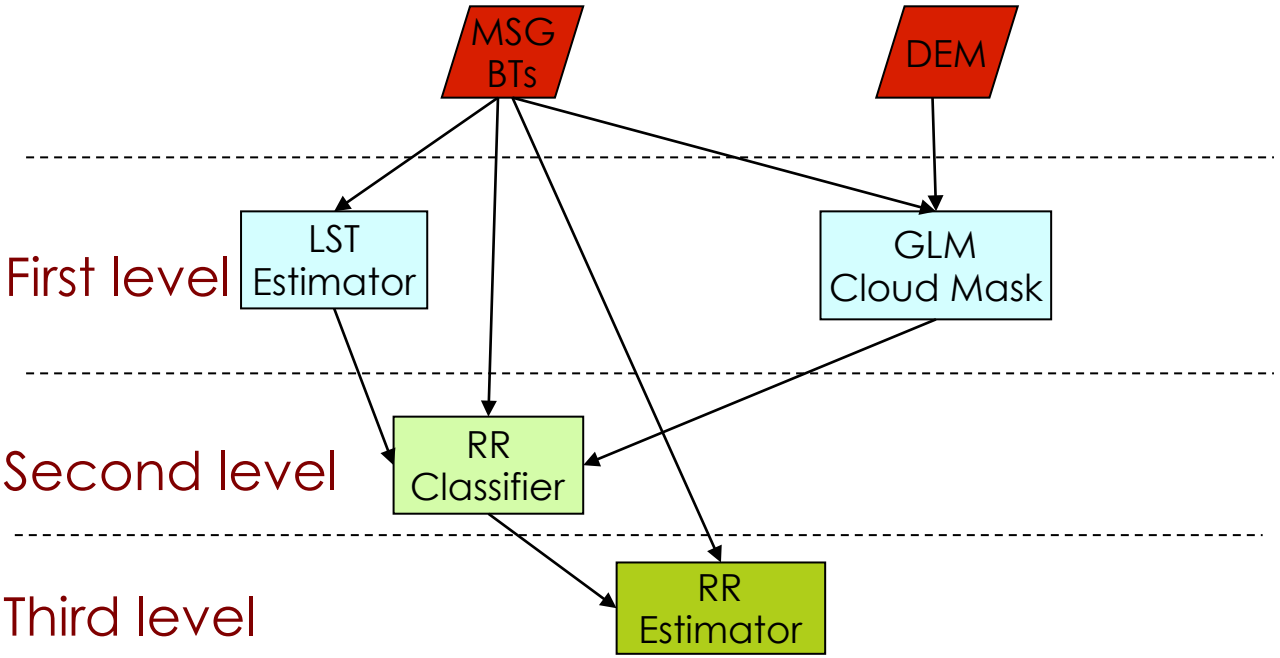
“Istant Weather” is a new app distributed on Google play for weather nowcasting. A neural network ensemble is applied to the data provided by Meteosat Second Generator system (Infrared channels) to predict the new satellite measurements and, from them, the evolution of geophysical fields of interest as:

- Cloud coverage
- Rain rate
- Temperature

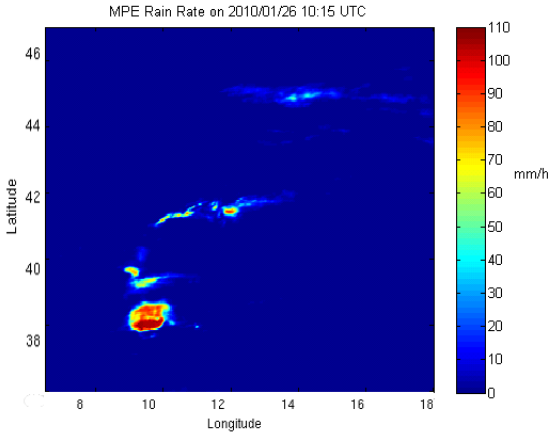
The Instant Weather app

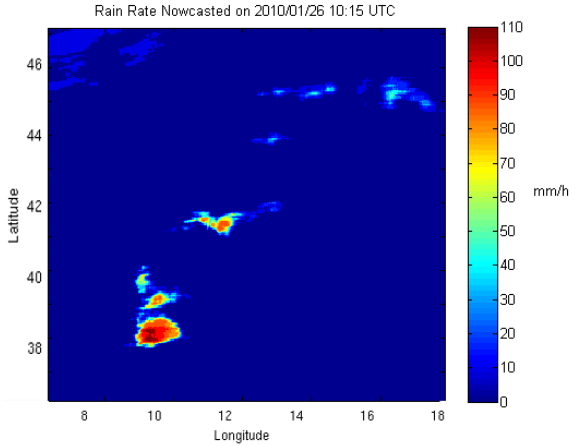
https://play.google.com/store/apps/details?id=org.mondometeo.instantweather

Estimations: the model layout



The rainfall estimation: a static case





Performance Indexes
60 Min

BIAS	**1.33**	**mm/h**
RMSE	**9.05**	**mm/h**
Correlation	**68.47**	**%**

Objectives

Although the applied techniques are very performing and reliable a complete characterization of the possible inaccuracies regarding rainfall and ground temperature values is still missing

The objectives are:

- to validate the results produced by the nowcasting system using ground truth data coming from many different sources
- to ensure final data to be of the highest possible quality and reliability

The Validation Process

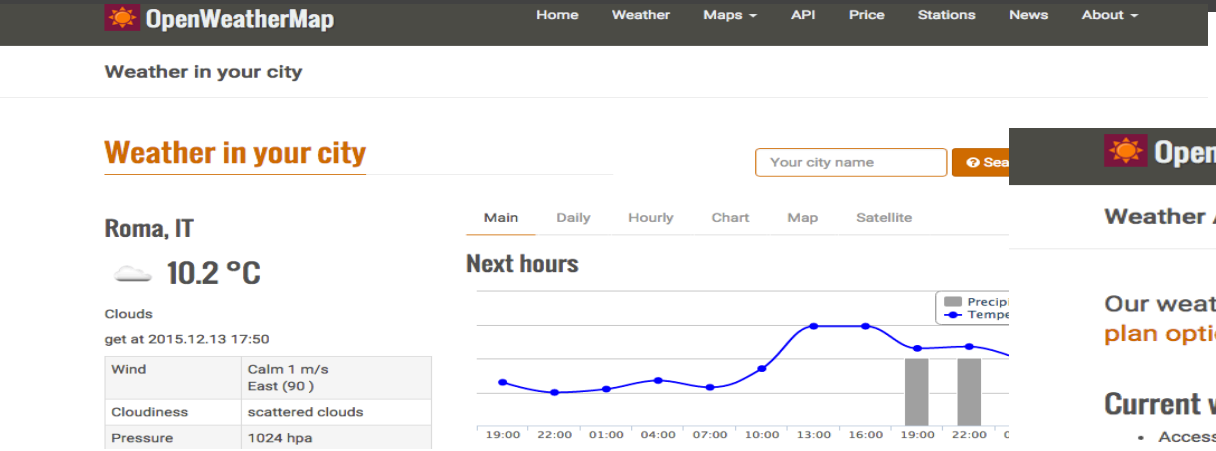
Validation by Ground Truth Data

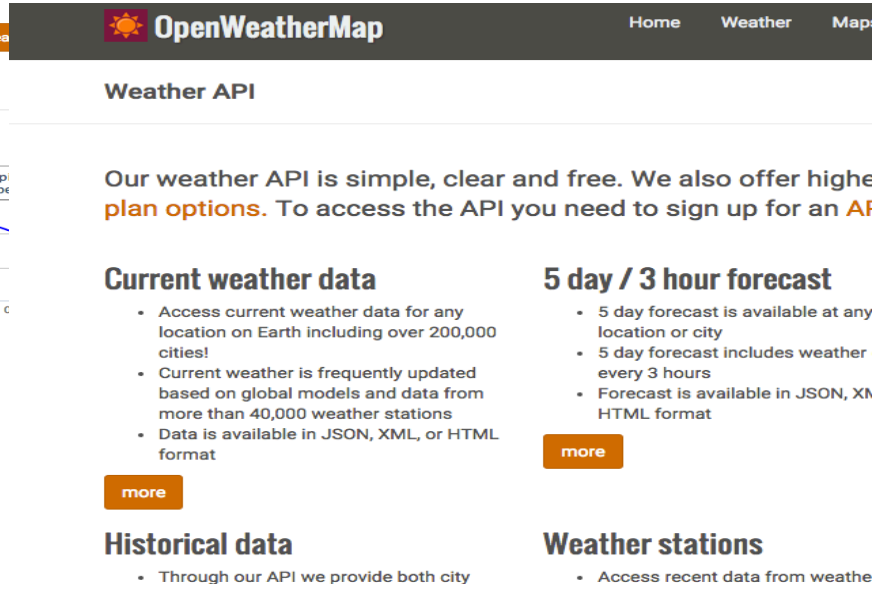
In the validation process data can be collected from various ground-based sources and sensors (in situ measurements, instruments, crowd-sourcing, open source platform), then quality-controlled, and finally compared with the satellite products in order to get validated retrievals.

Ground Truth data are collected from online and open data services:

- Openweathermap
- Wunderground
- Open Meteo Foundation

Weather API





They provide free API (Application Programming Interface) to weather data including current weather data, forecasts and history data.

Weather API

INPUT

api.openweathermap.org/data/2.5/weather?lat=35&lon=139

OUTPUT

{"coord":{"lon":139,"lat":35},

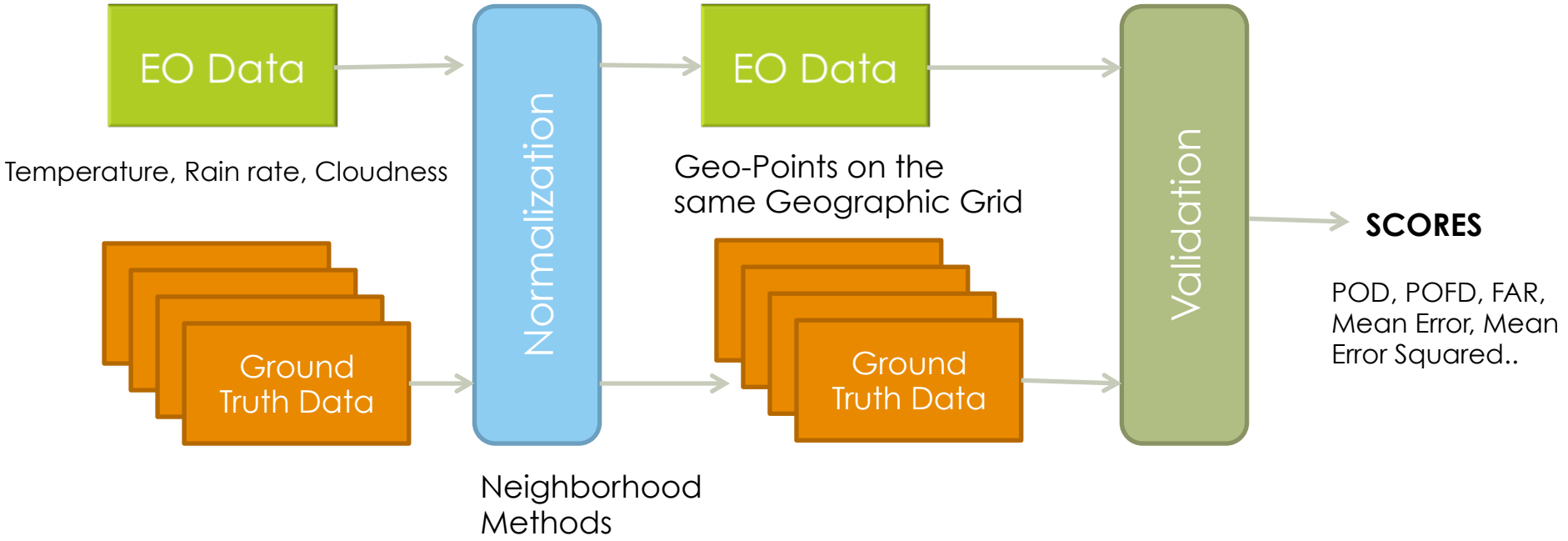
"sys":{"country":"JP","sunrise":1369769524,"sunset":1369821049},

"weather":[{"id":804,"main":"clouds","description":"overcast clouds","icon":"04n"}],

"main":{"temp":289.5,"humidity":89,"pressure":1013,"temp_min":287.04,"temp_max":292.04},"wind":{"speed":7.31,"deg":187.002},

"rain":{"3h":0},"clouds":{"all":92},"dt":1369824698,"id":1851632,"name":"Shuzenji","cod":200}

Validation Pipeline



Big Data issue #1 : Variety

“Variety” Critical Issues

- JSON responses coming from the different data sources are not formatted in the same way, no standards
- RDBMS (Relational Database): high data loading phase (too slow for Big Data), no flexibility to handle un-structured or semi-structured data
- Need of schema-less dabases

A possible solution: MongoDB

- It is a Non-Relational DB
- It supports Schema-less o No-Schema design
- It is “Document Oriented”
- It works with JSON (Javascript Object Notation)

A possible solution: MongoDB

- MongoDB stores data as “documents”, structured as JSON files and documents as “collections”
- No matters if documents have NOT the same structure!!
- MongoDB is a “schema-less” system: inserted docs have often different schemas

```
{
    _id: ObjectId("5099803df3f4948bd2f98391"),
    name: { first: "Alan", last: "Turing" },
    birth: new Date('Jun 23, 1912'),
    death: new Date('Jun 07, 1954'),
    contribs: [ "Turing machine", "Turing test", "Turingery" ],
    views : NumberLong(1250000)
}
```

Conclusion

MongoDB is the perfect choice for Ground Truth Data Storage, because:

- It can support the different JSON data file coming from the N sources
- It is scalable and flexible
- It is free ☺

Big Data issue #2 : Volume

Major complexities for data process

- Size of the images (>100Mb)
- $>10^5$ satellite pixels

complicate image handling including storing, and processing for enhancing the resolution making direct application of existing methods not possible

Proposed Framework

- The proposed framework relies on a parallel processing, Map Reduce system for data processing environment in order to process this big-data in reasonable time as in typical big data approaches
- In particular the proposed framework relies on the Hadoop framework for providing such features
- It is a highly scalable model for distributed programming on clusters of computer (created by Google)

Map Reduce Model

- It relieves the burden of the programmers dealing with distributed programming, since it hides the details of parallelization, fault-tolerance, locality optimization, and load balancing
- A wide range of computing problems could be presented in MapReduce model, e.g. sorting, data mining, machine learning, image processing, and many other systems;
- The scability of MapReduce is up to thousands of machines which is suitable for the real workload.

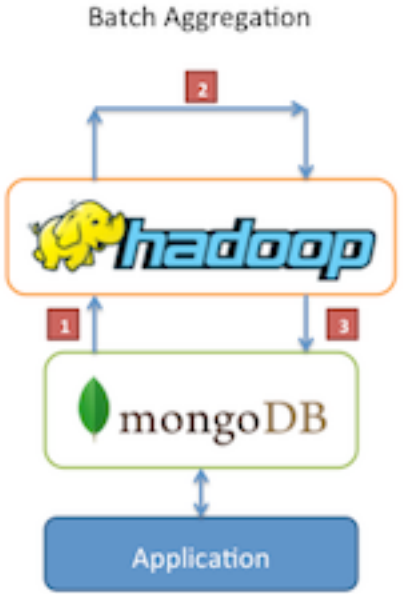
Map Reduce: how it works

- MapReduce is a programming model designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks
- It works by breaking the processing in two phases: the **MAP** phase and the **REDUCE** phase
- Data are colocated with compute nodes. Data access is LOCAL!!
- Hadoop is an implementation of the “map-reduce” programming model and is designed to scale up from single server to thousands of machines, each offering local computation and storage

The Proposed Solution

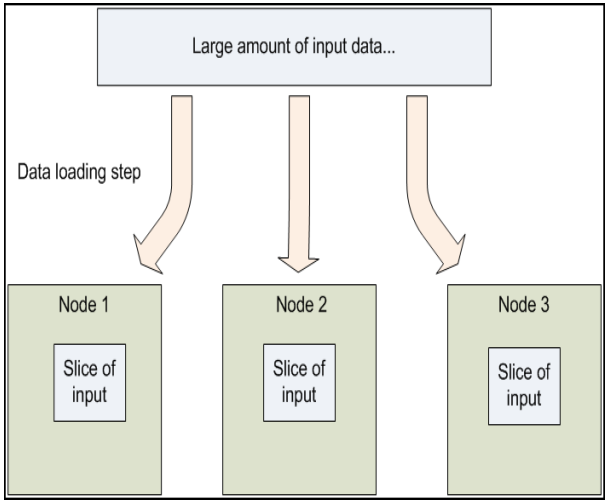
Data Acquisition

- Ground Truth Data and Instant Weather Data are stored into MongoDb collections
- Data is pulled from MongoDB and processed within Hadoop via one or more MapReduce jobs
- Output from these MapReduce jobs can then be written back to MongoDB for later querying and ad-hoc analysis



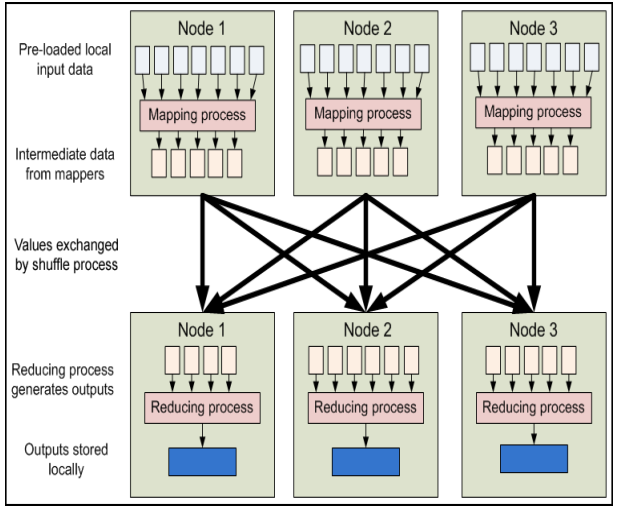
Data Processing

Since data is spread across the distributed file system as chunks, each compute process running on a node operates on a subset of the data. This strategy of moving computation to the data, instead of moving the data to the computation allows Hadoop to achieve high data locality which in turn results in high performance.

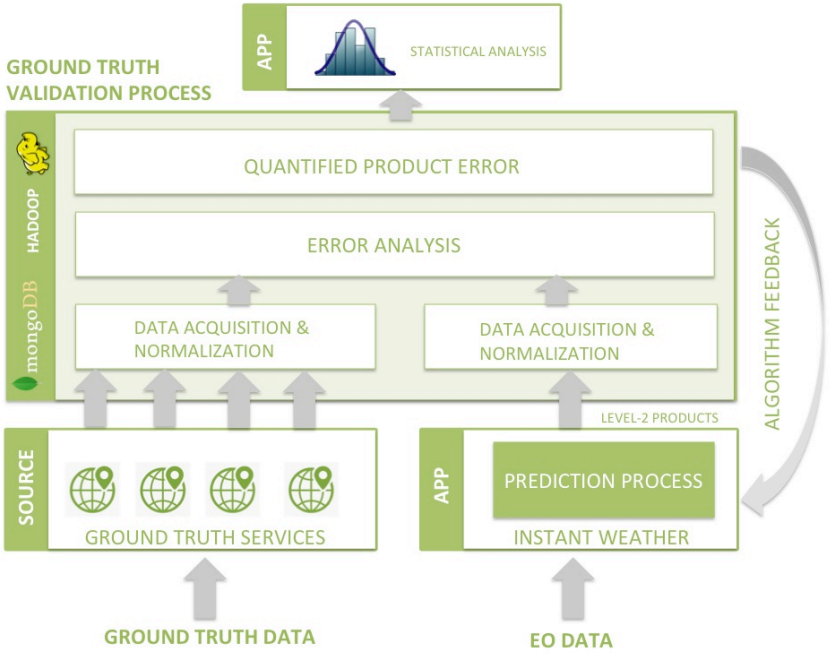


Data Processing

In MapReduce, records are processed in isolation by tasks called Mappers. The output from the Mappers is then brought together into a second set of tasks called Reducers, where results from different mappers can be merged together



Final Architecture



Conclusions

The methodology ensures a validation process on EO satellite data and provides a high quality weather nowcasting service to the end users.

Furthermore this methodology, based on open data and an open source solution as Hadoop opens up great opportunities also to small and medium companies to play on large scaled system that support multiple terabytes and load thousands of transactions per second, performing complex transformations and analysis on massive amounts of data in short time.

Thanks for your attention

giuseppe.iannitto@gmail.com