Big Data Storage for Hospital Drugs Prescriptions and Parkinson Disease Risk Detection

Rodrigo Sestari, Charles Ferrari
University of Trento

Scenario

- **Our Goal:** Create a Big data source for hospital drugs administrations and perform data analysis for Parkinson risk detection.
- *Medicine administration*
  - Prescription > Ordering > Dispensing > Monitoring
- *Data Source*
  - 149.490 drugs administrations from 31st of July 2015 to 30th of November 2015
  - 4944 different therapies (each therapy is one inpatient admission)
  - Inpatient information like Age, Sex and admission
  - Drug information like ATC Code and Health Ministry code.
  - Data from the hospital's logs of medicine administrations that is produced by the medicine's trolleys, with information about the drug and the inpatient.
  - Data about presence of lactose or glucose on the medicine
  - Data from different sources

| Code	Content
A	Alimentary tract and metabolism
B	Blood and blood forming organs
C	Cardiovascular system
D	Dermatologicals
G	Genito-urinary system and sex hormones
H	Systemic hormonal preparations, excluding sex hormones and insulins
J	Antiinfectives for systemic use
L	Antineoplastic and immunomodulating agents
M	Musculo-skeletal system
N	Nervous system
P	Antiparasitic products, insecticides and repellents
R	Respiratory system
S	Sensory organs
V	Various

Main Data Provider

- **Drugs Trolley:** Every time a medicine is administered into a hospital's room, the system give information to a central JBoss server.

Performance Tests

- **First test** is represented by the red arrow: on this test we sent a simple log file with 2Mb of data.
- **Second test**, represented by the blue arrow: We sent eight different log files, each one containing 52.2 megabytes sent in 50 seconds,
- **Third test** represented by the black arrow: two log files of 52.2 megabytes of data.
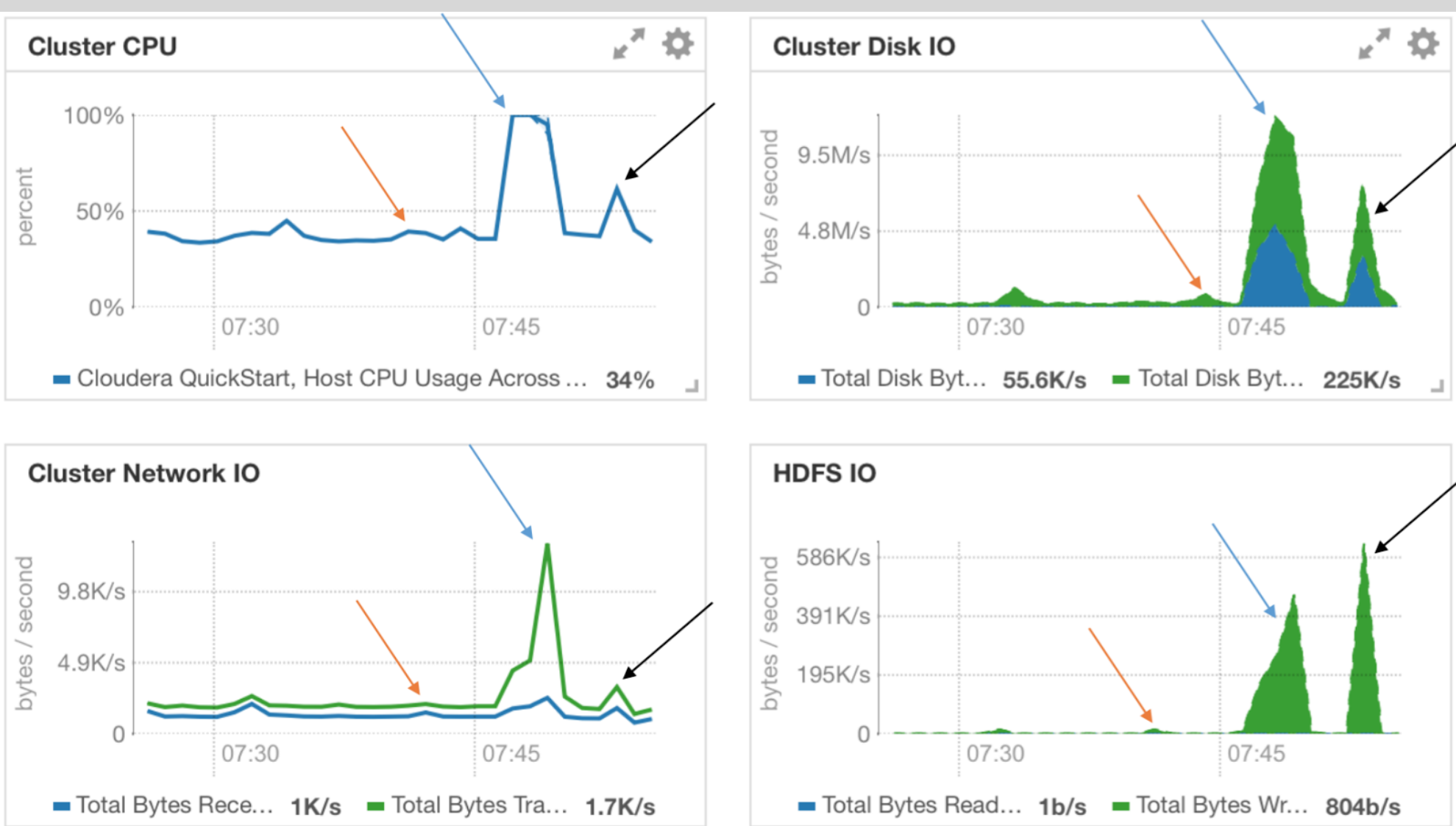
Solution

Overview:

- **Data Acquisition:** Collect data from the Trolley's server with Apache Kafka Message Broker.
- *Data Parsing using Spark Streaming:* Data may be sent twice and it has several information that are not relevant for us, for that, we clean the data and parse it removing everything that does not concern inpatients or the medicines.
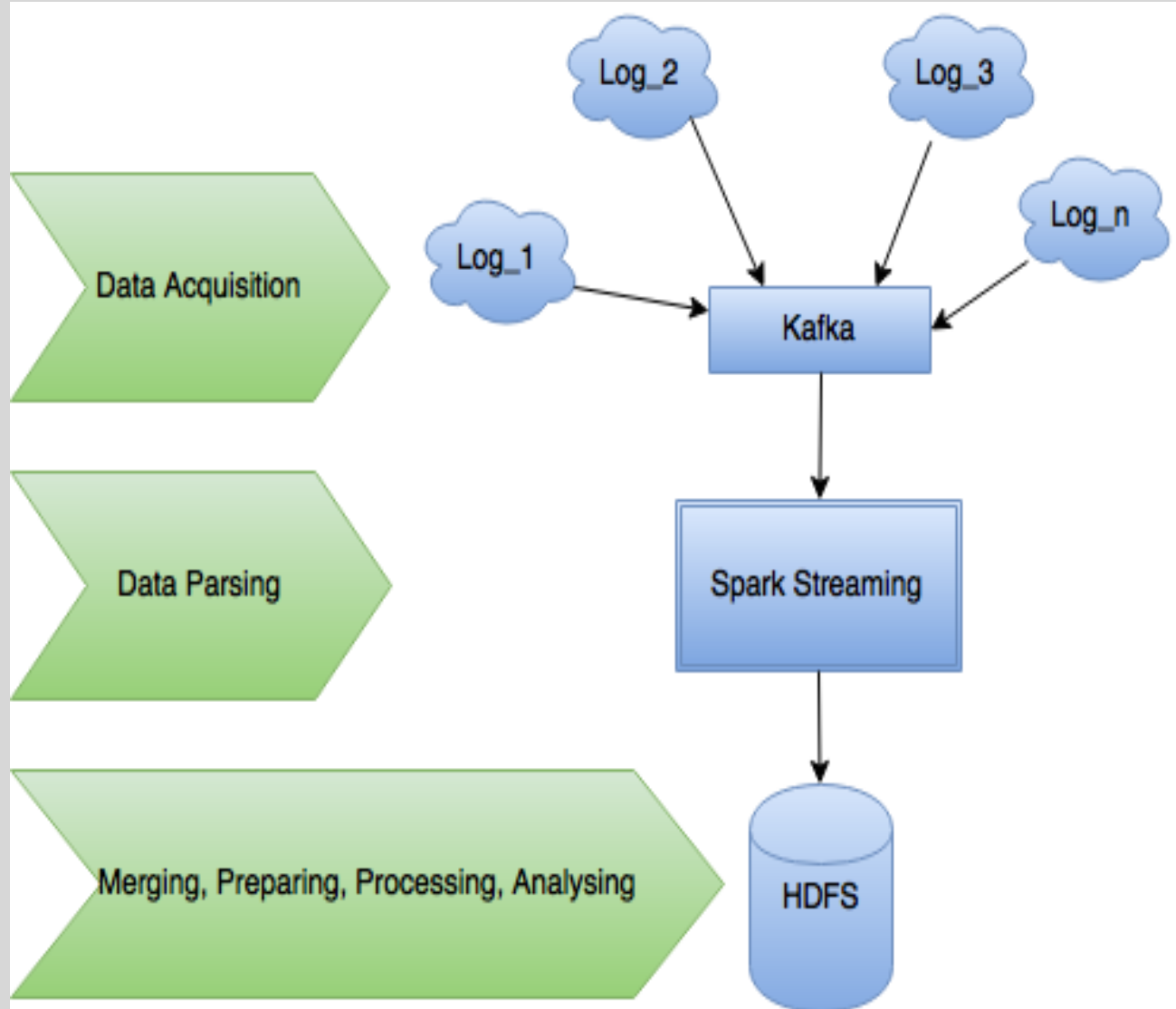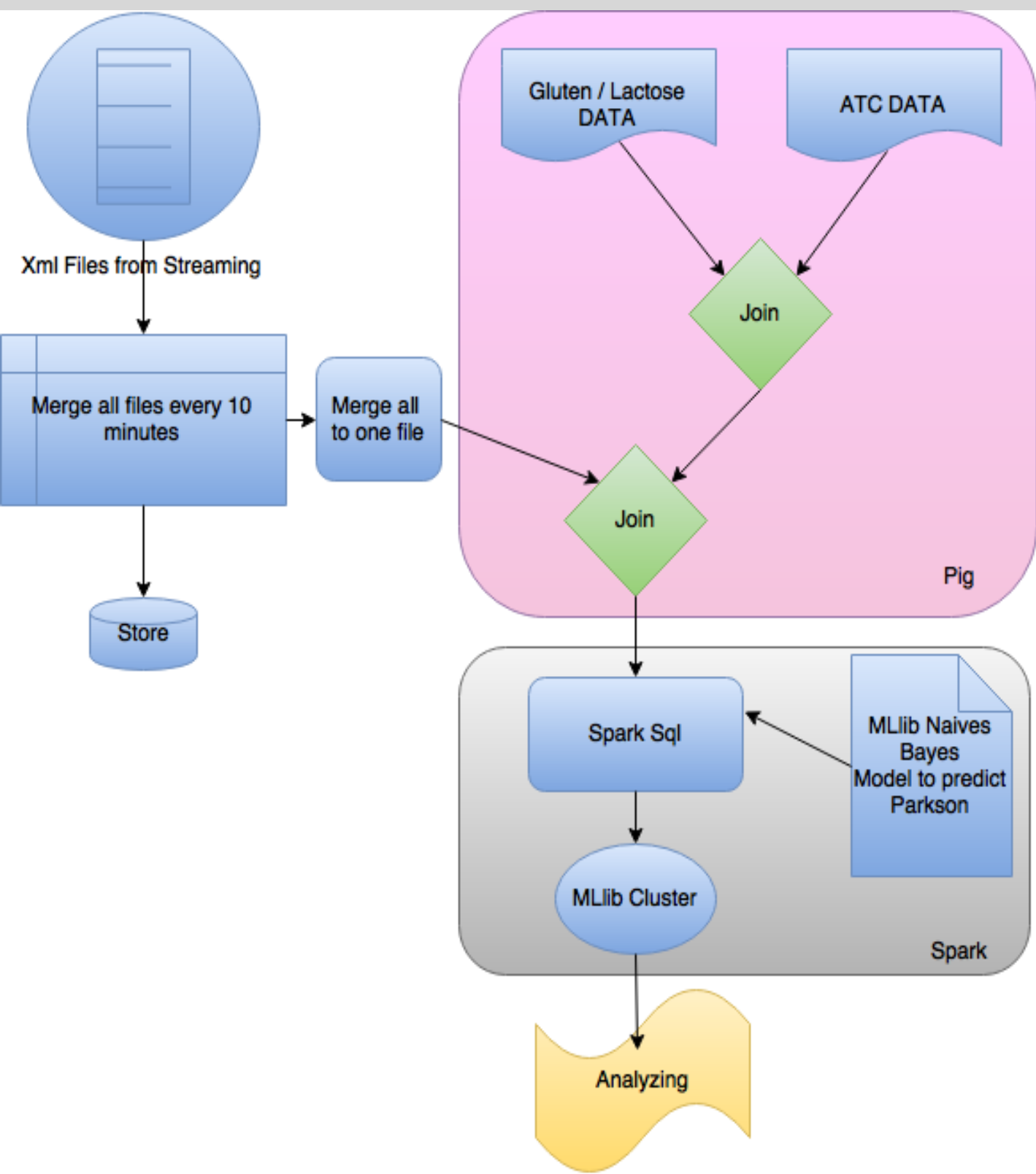
Data Merging and Preparing:

On this step we want to merge some other information and prepare our data to be saved and sent to the Parkinson Risk analyze

- *Merging data from other sources:* Merging information regarding presence of lactose and glucose. Merging ATC Code information instead of the Health Ministry code.
- *Preparing data*. Identifiing the inpatients and grouping the data per patient with its respective ATC Code and identifying drugs that may represent risk of Parkinson disease.

Clustering according to Parkinson development risk

With the given data we will process it into clusters and save them.
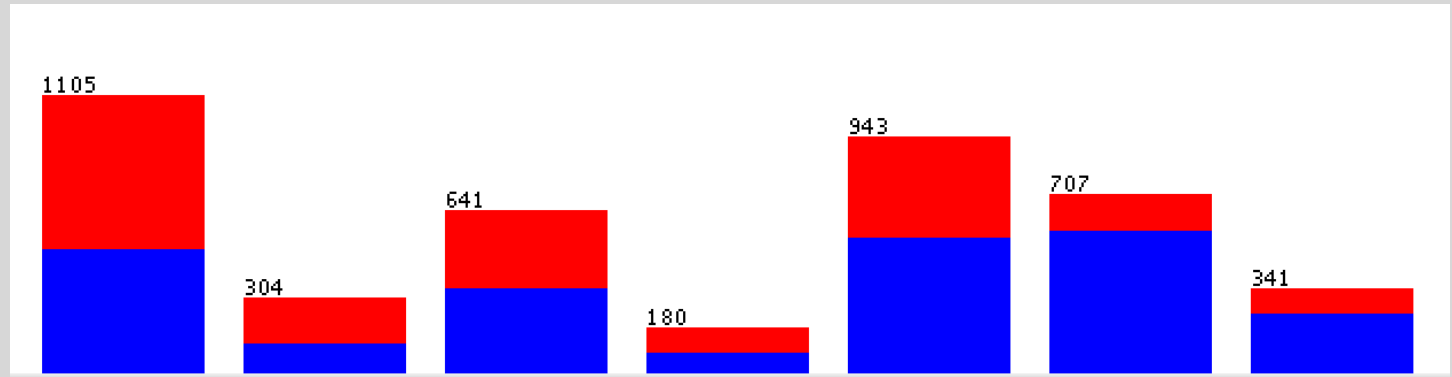
- *Creating Clusters:* Following recent studies and using Naive Bayes for classifying, we selected the medicines that its use can sign a risk of Parkinson disease development. Seven clusters were created, see top figure on the right.
- *Data Analyze:* The clusters are represented on the right, Blue represents zero percent, red represents five percent, green represents twenty percent and the orange points located on the cluster predominantly green, represents 50% of risk