Smart XML Platform

XML files ingestion and calculation powered by Cloudera Ecosystem

SFERANET

Introduction

- There is plenty of tools for handling data in JSON
- Many companies still keep their data in XML format
- In XML you have a default way to validate against a schema

Business use case

- Today companies have terabytes of data
- Usually stored in XML files
- Files have to be validated against an XSD schema
- We also need a logical validation: is the information coherent with other data?

Smart XML Platform

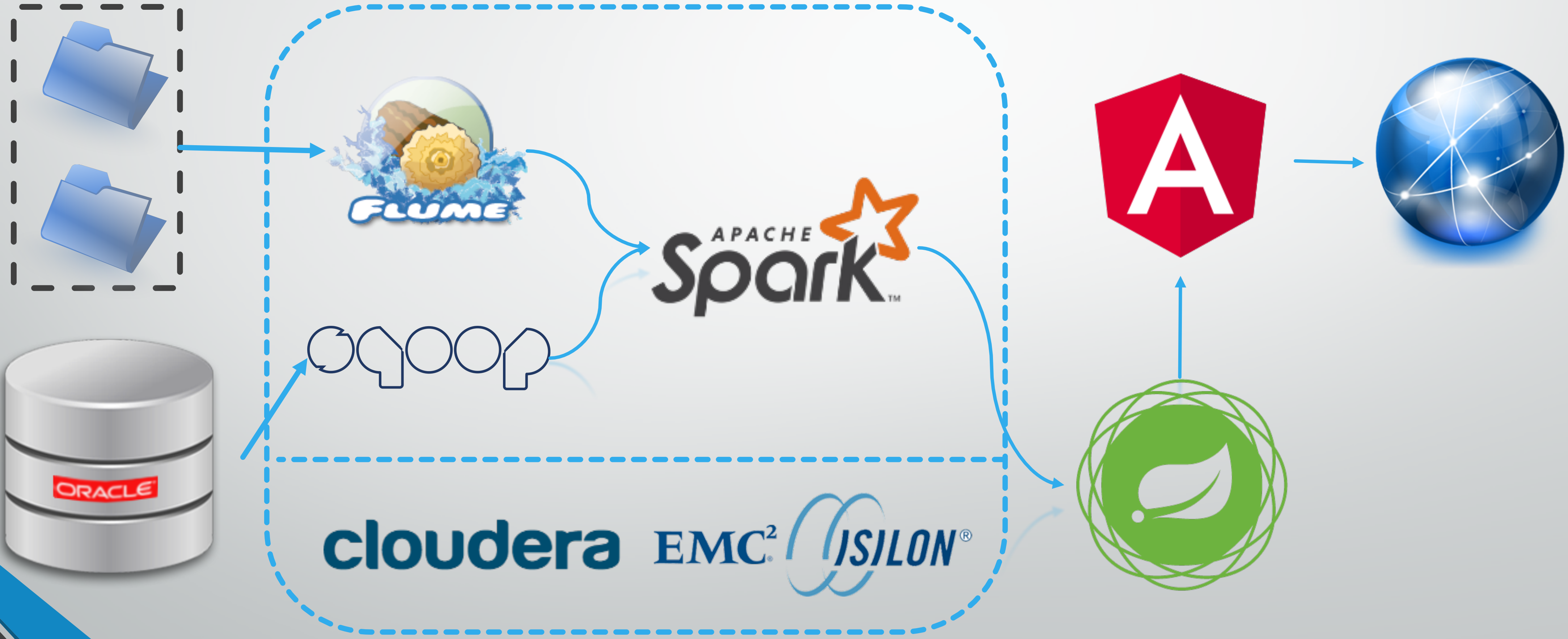
Three steps:

- Reading the data
- Perform business logic
  - Validation
  - Aggregation
- Write the result on HDFS

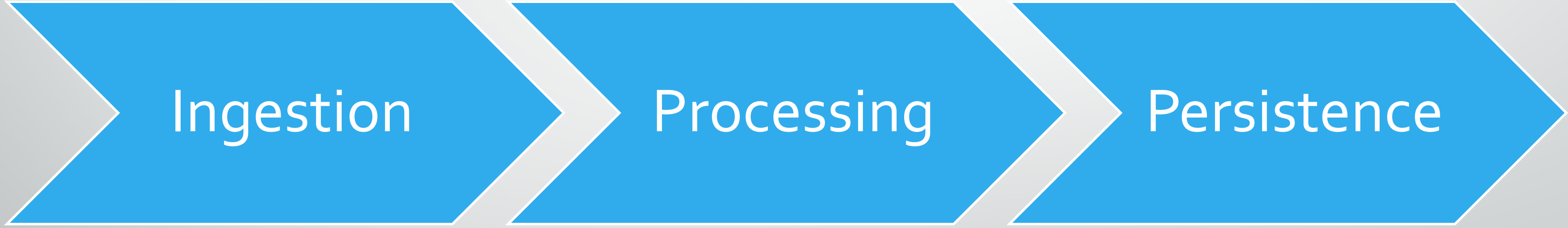
We made it FAST

- It runs on Cloudera Distribution including Hadoop (CDH)
- Archive files are partitioned on HDFS
- Read and analyzed with Apache Spark

Ingestion and Calculation Flow



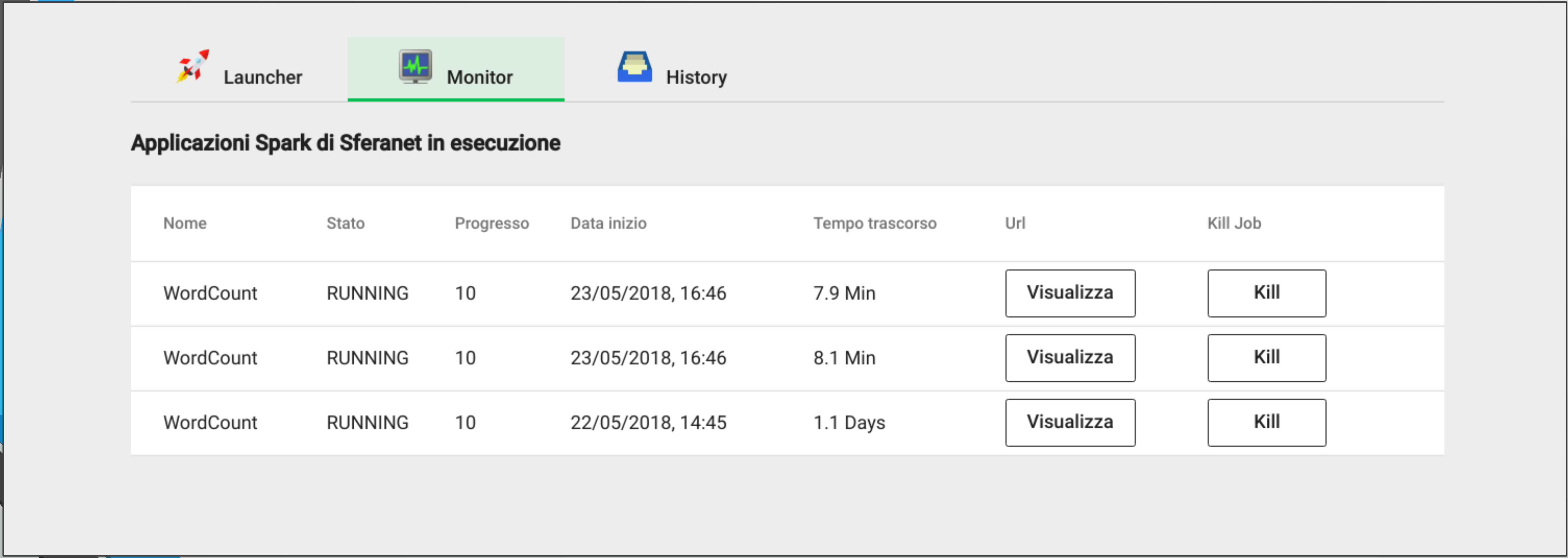
Workflow



A flexible solution

- Provide the XML Schema Definition File (XSD)
- Smart XML Engine will analyze data formats
- Specify additional business constraints using JSON
- Specify processing operations via another JSON file
- Ready to go!

The dashboard



Applicazioni Spark di Sferanet in esecuzione

Nome	Stato	Progresso	Data inizio	Tempo trascorso	Url	Kill Job
WordCount	RUNNING	10	23/05/2018, 16:46	7.9 Min	Visualizza	Kill
WordCount	RUNNING	10	23/05/2018, 16:46	8.1 Min	Visualizza	Kill
WordCount	RUNNING	10	22/05/2018, 14:45	1.1 Days	Visualizza	Kill