Big Data Architectures & Real World Experiments

Journée Big Data & HPC - Grenoble, 26 avril 2016

Yann Vernaz

Outline

1. Introduction

2. Big Data Architectures

3. Real World Experiments

4. Summary

Introduction

ERODS Team (27 pers.)

Efficient and **RO**bust **D**istributed **S**ystems

Responsable : Noël De Palma

Equipe de recherche commune CNRS, Grenoble INP, UJF, UPMF

Site Web : http://erods.liglab.fr

Processing Models

Batch Processing

- Familiar concept of processing data en masse.
- Generally incurs a high-latency.

(Event-) Stream Processing

- A one-at-a-time processing model.
- A datum is processed as it arrives.
- Sub-second latency.
- Difficult to process state data efficiently.

Micro-Batching

- A special case of batch processing with very small batch sizes.
- A enjoyable mix between batching and streaming.
- At cost of latency.
- Gives stateful computation, making windowing an easy task.

Big Data Architectures

Requirements dictate the choice

Latency

Is performance of streaming application paramount.

Development Cost

Is it desired to have similar code bases for batch and stream processing.

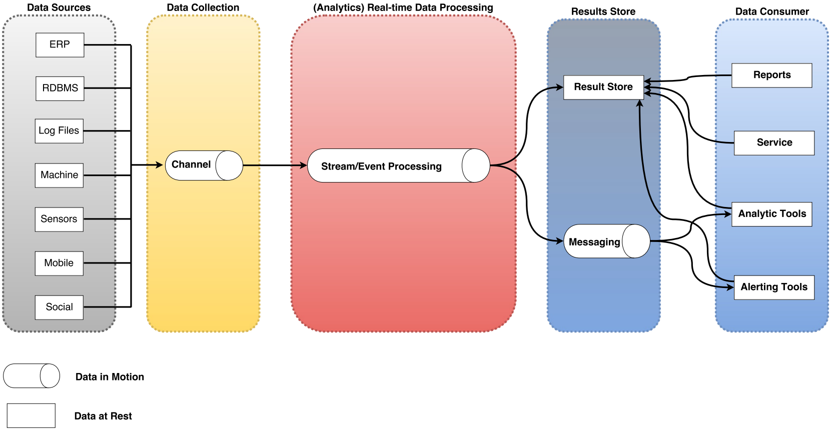
Message Delivery Guarantees

Is there high importance on processing every single record, or is some normal amount of data loss acceptable.

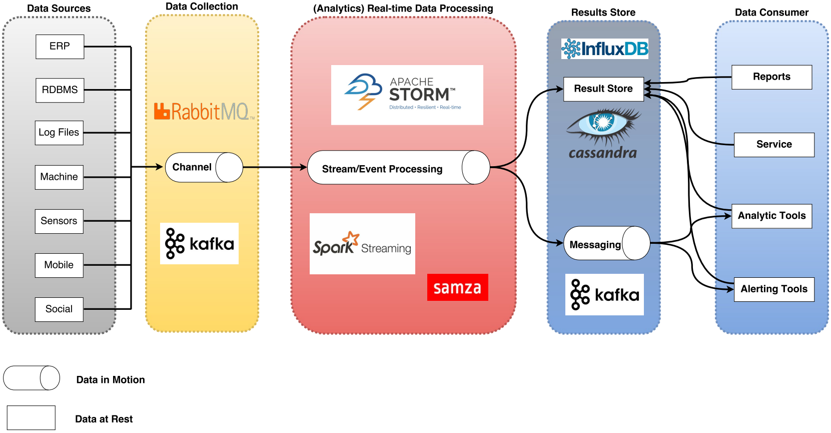
Process Fault Tolerance

Is high-availability of primary concern.

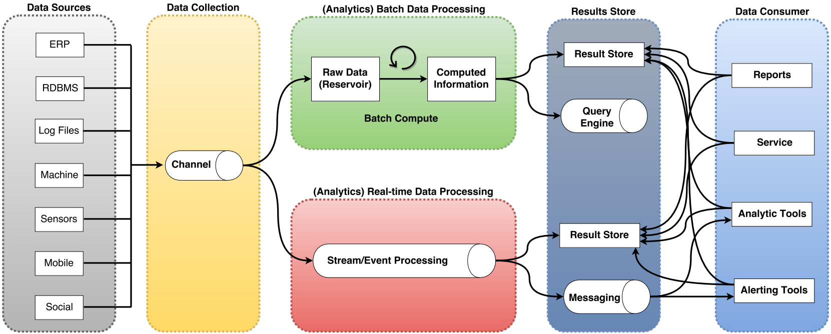
Streaming Analytics Architecture

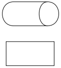


Streaming Analytics Architecture (Technologies)

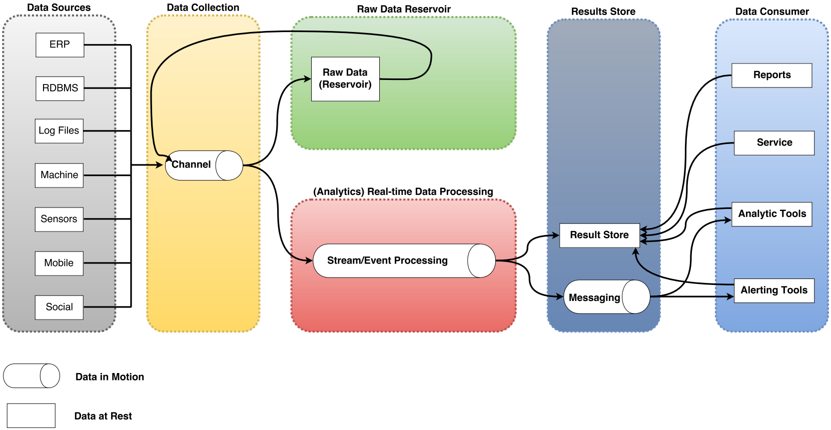


Lambda Architecture

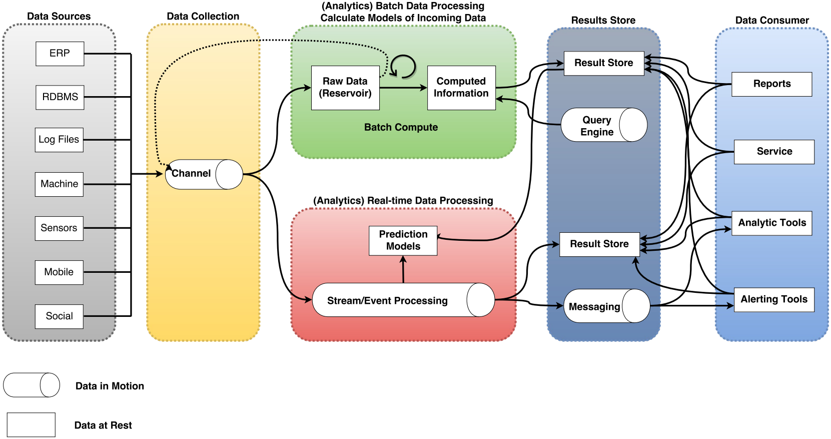




Kappa Architecture



Unified Architecture



Data Sources

ERP

RDBMS

Log Files

Machine

Sensors

Mobile

Social

Data Collection

Channel

(Analytics) Batch Data Processing
Calculate Models of Incoming Data

Raw Data (Reservoir)

Computed Information

Batch Compute

(Analytics) Real-time Data Processing

Prediction Models

Stream/Event Processing

Results Store

Result Store

Query Engine

Result Store

Messaging

Data Consumer

Reports

Service

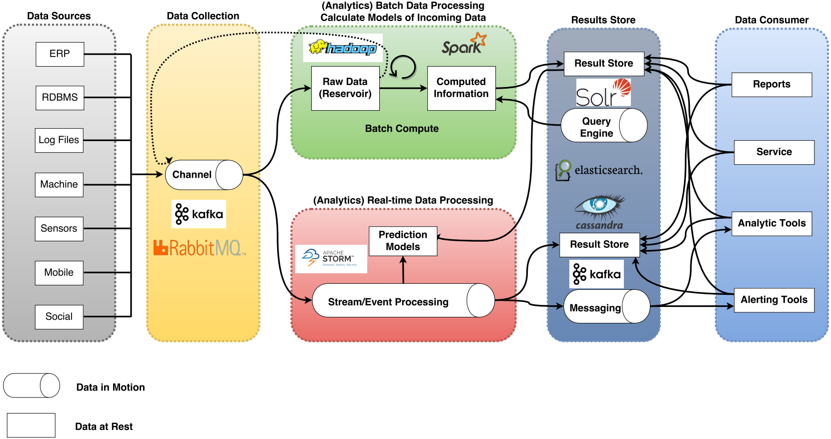
Analytic Tools

Alerting Tools

Data in Motion

Data at Rest

Unified Architecture (Technologies)

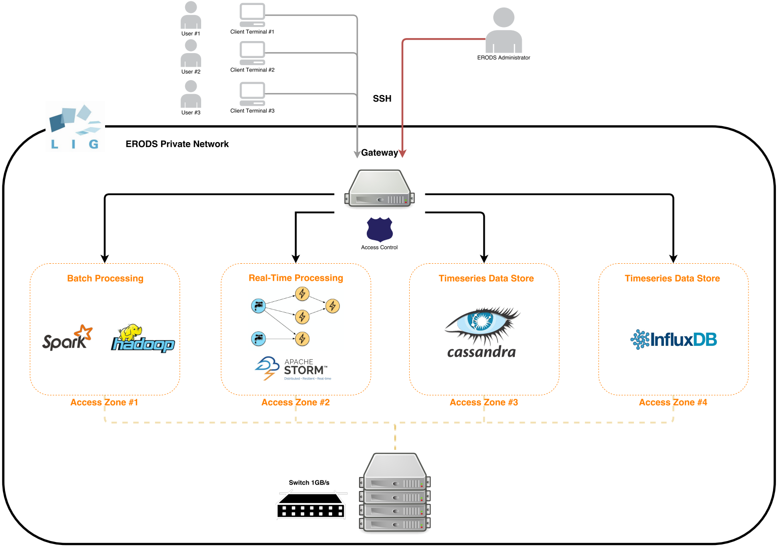


Real World Experiments
Proof-Of-Concept

Virtualization: OpenStack + KVM

- Our experimental platform is Openstack-based.
- OpenStack has been deployed using Fuel.

Services Architecture



User #1

Client Terminal #1

User #2

Client Terminal #2

User #3

Client Terminal #3

ERODS Administrator

SSH

ERODS Private Network

Gateway

Access Control

Batch Processing

Access Zone #1

Real-Time Processing

Access Zone #2

Timeseries Data Store

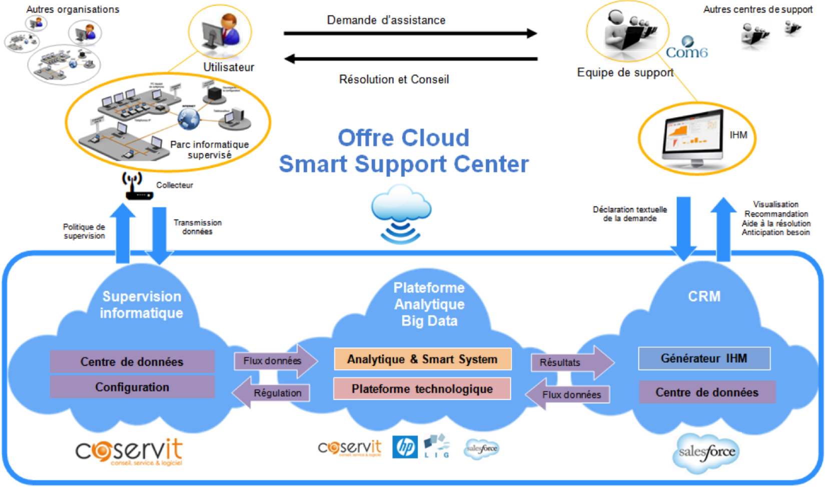
Access Zone #3

Timeseries Data Store

Access Zone #4

Switch 1GB/s

Smart Support Center - Description



Autres organisations

Utilisateur

Parc informatique supervisé

Collecteur

Demande d'assistance

Résolution et Conseil

Equipe de support

Com6

Autres centres de support

IHM

Offre Cloud
Smart Support Center

Politique de supervision

Transmission données

Déclaration textuelle de la demande

Visualisation
Recommandation
Aide à la résolution
Anticipation besoin

Supervision informatique

- Centre de données
- Configuration

Flux données

Régulation

Plateforme Analytique Big Data

- Analytique & Smart System
- Plateforme technologique

Résultats

Flux données

CRM

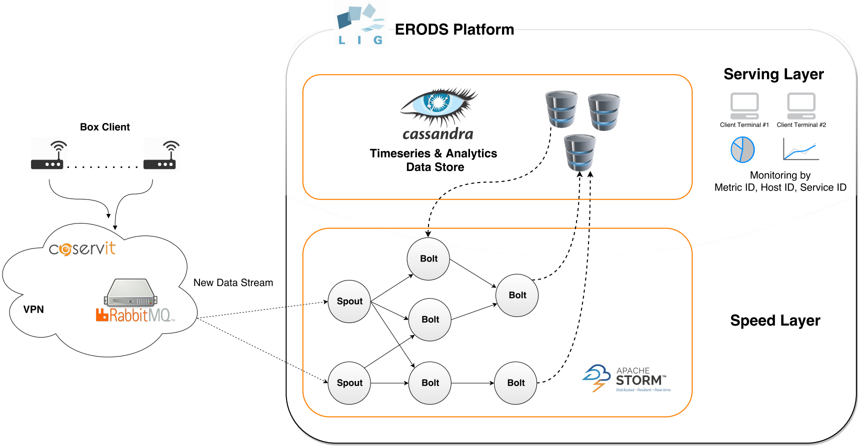
- Générateur IHM
- Centre de données

coservit conseil, service & logiciel

coservit hp LIG salesforce

salesforce

Smart Support Center - Streaming Analytics



Box Client

coservit

VPN

RabbitMQ

New Data Stream

L I G

ERODS Platform

cassandra

Timeseries & Analytics Data Store

Serving Layer

Client Terminal #1

Client Terminal #2

Monitoring by Metric ID, Host ID, Service ID

Spout

Spout

Bolt

Bolt

Bolt

Bolt

Bolt

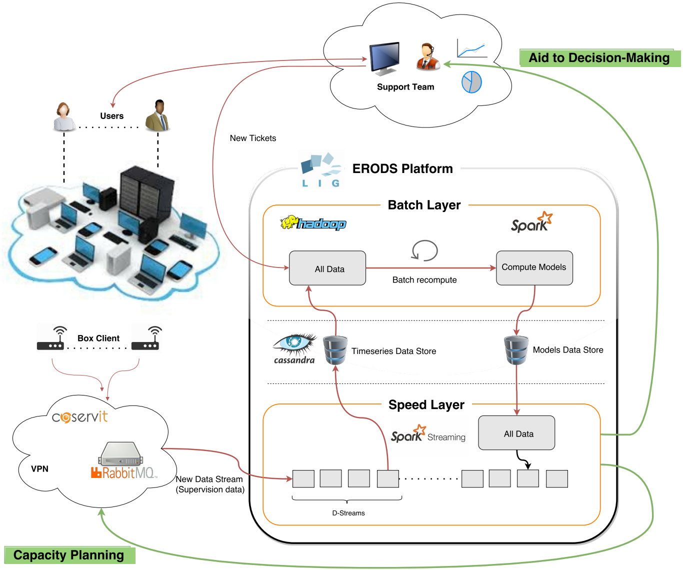
Bolt

APACHE STORM™

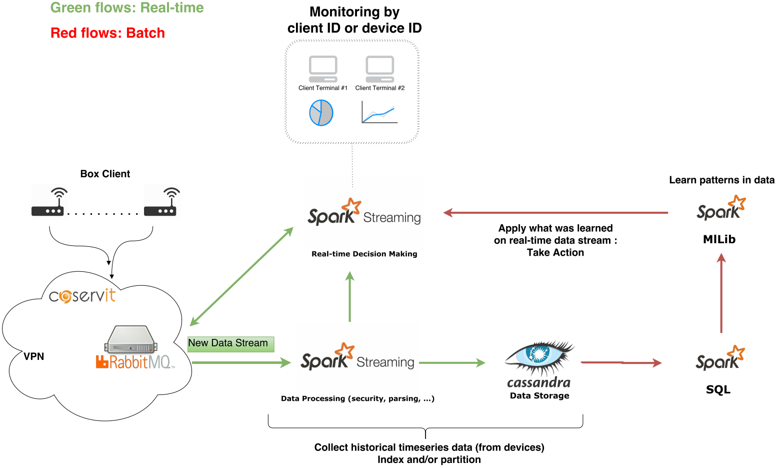
Distributed • Resilient • Real-time

Speed Layer

Smart Support Center - Analytics System



Smart Support Center - Data Flow



Green flows: Real-time

Red flows: Batch

Monitoring by
client ID or device ID

Client Terminal #1

Client Terminal #2

Box Client

VPN

New Data Stream

Spark Streaming

Real-time Decision Making

Spark Streaming

Data Processing (security, parsing, ...)

cassandra
Data Storage

Spark SQL

Spark MlLib

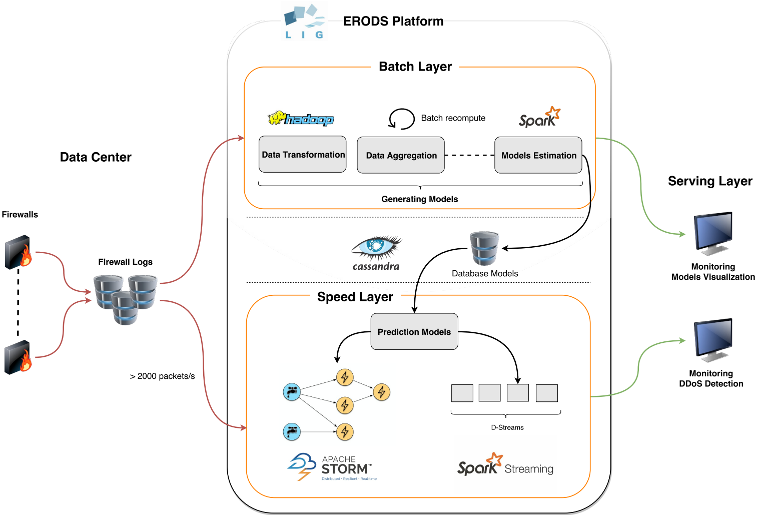
Learn patterns in data

Apply what was learned
on real-time data stream :
Take Action

Collect historical timeseries data (from devices)
Index and/or partition

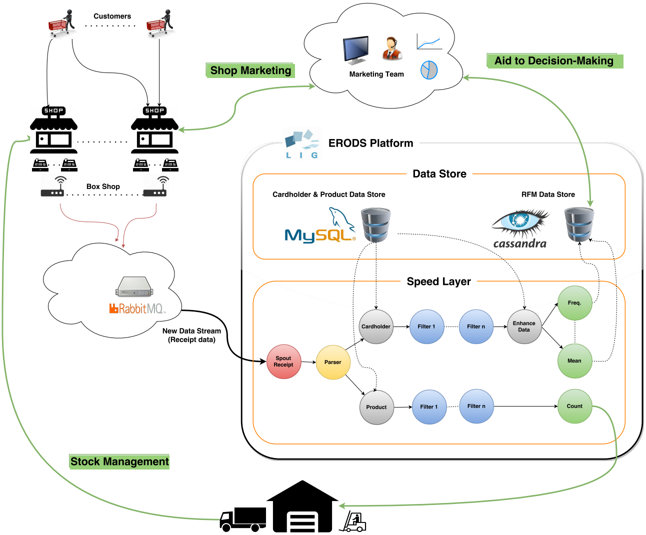
DDoS Attack Detection - Description

DDoS Attack Detection - Architecture



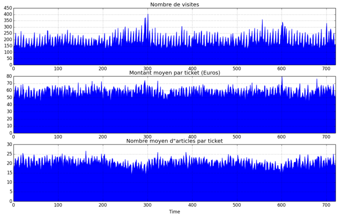
Churn Risk Analysis

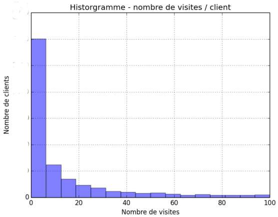
Churn Risk Analysis - Streaming Analytics

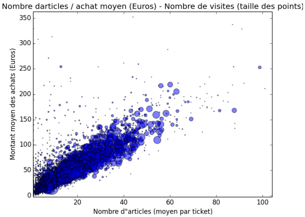


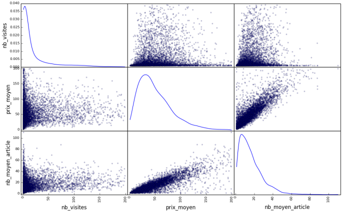
Compute real-time statistics by client, shop or country.

Churn Risk Analysis - Real Time Reporting

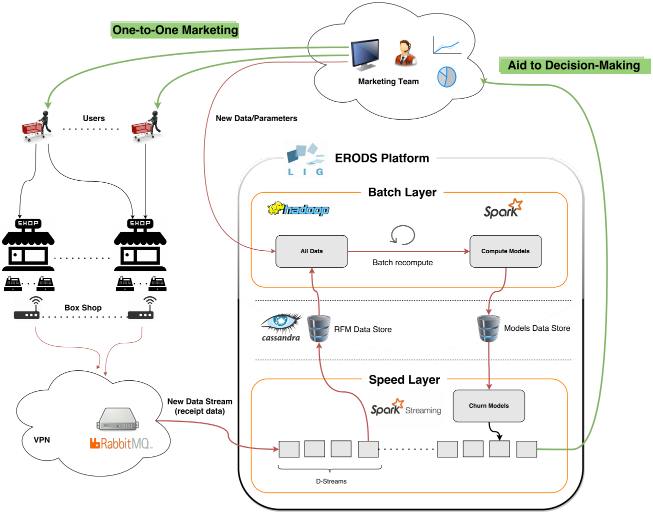






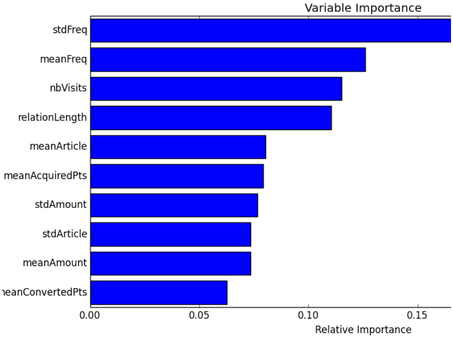


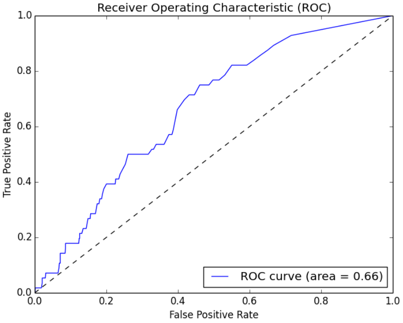
Churn Risk Analysis - Customer Churn Probability



Compute real-time churn risk probability by client.

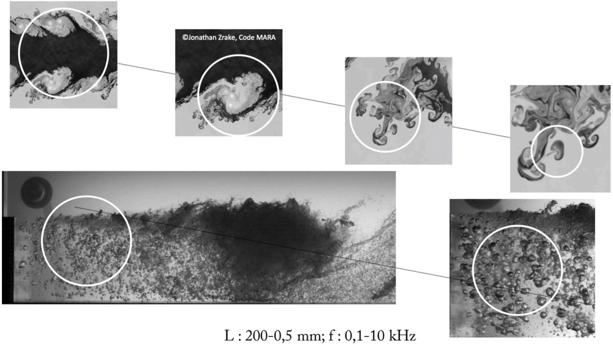
Churn Risk Analysis - Results.





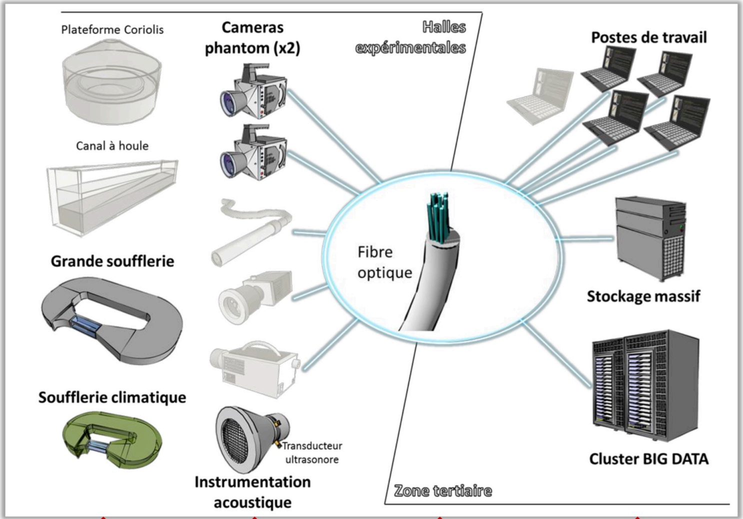
Projet STREAM

Objet de l'étude : écoulements turbulents et/ou multiphasique.



L : 200-0,5 mm; f : 0,1-10 kHz

Projet STREAM



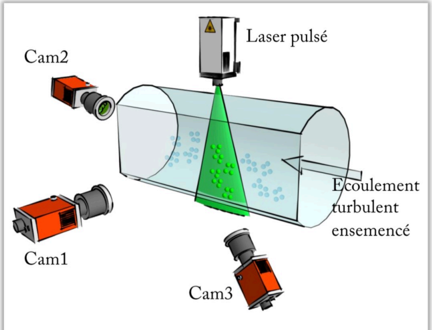
Adaptation des installations

Nouvelle instrumentation

Infrastructure fibre optique

Nouvel environnement Big Data

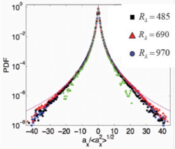
Projet STREAM



Dispositif expérimental

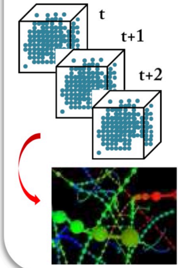
L : 5 cm-50 µm; t : 1-100 ms (eau)

Mesure d'accélération : 100 000 i/s



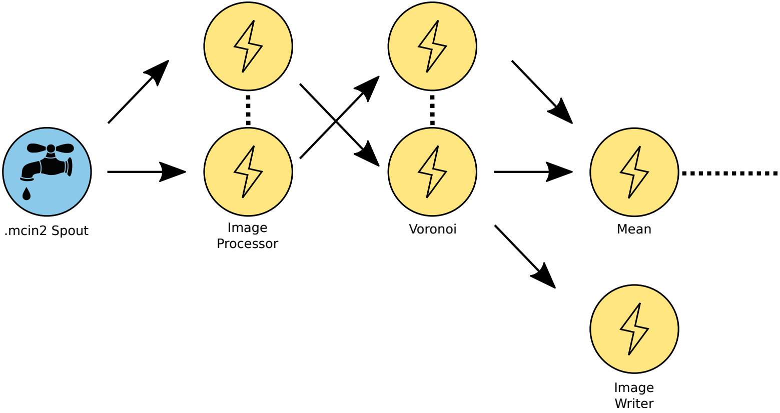
Suivi lagrangien

Suivi dans le temps de chaque particule



- 3x24 GB toutes les 20 s soit **300 TB/jour** d'images brutes...
- Extraction des positions en temps réel.
- Gestion des tailles, trajectoires et vitesses des particules.

Projet STREAM - Topology Storm



Summary

Big Data is still "work in progress"

Choosing the right architecture is key for any (big data) project.

Know the use cases before choosing your architecture.

There are no standard architectures available which have been used for years (young field).

In the past few years, a few architectures have evolved and have been discussed.

To have one/a few reference architectures can help in choosing the right components.

From Offline to Online

Recently, Big Data analytics has been transitioning from **offline** (or batch) to primarily **online** (or streaming).

Forrester remarked the following in Q3 2014

"The high velocity, white-water flow of data from innumerable real-time data sources such as market data, Internet of Things, mobile, sensors, click-stream, and even transactions remain largely unnavigated by most firms. The opportunity to leverage streaming analytics has never been greater."

Velocity is one of the 7Vs commonly used to characterize Big Data.