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 RObust Distributed Systems

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 (Technologies)
Lambda Architecture
Unified Architecture (Technologies)
Real World Experiments
Proof-Of-Concept
Virtualization: OpenStack + KVM

- Our experimental platform is Openstack-based.
- OpenStack has been deployed using Fuel.
Smart Support Center - Description

[Diagram of Smart Support Center with components and processes]
Smart Support Center - Streaming Analytics

[Diagram showing a network architecture with components such as Box Client, VPN, RabbitMQ, cassandra Timeseries & Analytics Data Store, ERODS Platform, and Apache Storm.]

Serving Layer
- Monitoring
- Metric ID, Host ID, Service ID

Speed Layer
- Apache Storm
- Spout, Bolt components

EROODS Platform
- Data flow management

New Data Stream
DDoS Attack Detection - Description
DDoS Attack Detection - Architecture

Data Center

Firewalls

> 2000 packets/s

Firewall Logs

Batch Layer

Data Transformation
Data Aggregation
Models Estimation
Generating Models

EROODS Platform

Batch recompute

Spark

Speed Layer

Prediction Models

APACHE STORM

D-Streams

Spark Streaming

Serving Layer

Monitoring Models Visualization

Monitoring DDoS Detection

Cassandra

Database Models
Churn Risk Analysis
Churn Risk Analysis - Streaming Analytics

Compute real-time statistics by client, shop or country.
Churn Risk Analysis - Real Time Reporting
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

- Plateforme Coriolis
- Canal à houle
- Grande soufflerie
- Soufflerie climatique
- Cameras phantom (x2)
- Halles expérimentales
- Postes de travail
- Stockage massif
- Cluster BIG DATA
- Fibre optique
- Zone tertiaire
- Instrumentation acoustique

- Adaptation des installations
- Nouvelle instrumentation
- Infrastructure fibre optique
- Nouvel environnement Big Data
Projet STREAM

- 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.
Summary
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.
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.