ALTIC Big Data Stack

Charly Clairmont, ALTIC

@egwada
charly.clairmont@altic.org

http://www.altic.org
Our historical tools

- ETL: Talend
- Reporting: JasperReports, Birt
- OLAP: Mondrian, Palo
- BI platform: SpagoBI
Smart assembling
Innovation & customers' needs

• Identify when applied research is an opportunity for us, our solutions and our customers.

• Understand the business process of our customer & assess the impact of Open IT on their activities

• Offer an approach of the project both a technical and a operative

➔ Altic projects

➔ Allows our customer to optimize their business process

➔ Takes the customer job into account

➔ Offers perennial solutions

➔ Follows the customer present needs and not the editors' agenda
Identify Big Data potential / Hadoop

---

**hadoop-common-dev mailing list archives**

<table>
<thead>
<tr>
<th>Message view</th>
<th>From</th>
<th>Subject</th>
<th>Date</th>
</tr>
</thead>
</table>

```
41. Would love to see more such projects' integration with hadoop.

-Milind

----- Original Message ----- 
From: Ian Holoman <holoman.net>
To: core-dev@hadoop.apache.org <core-dev@hadoop.apache.org>

I'd like to volunteer a proposal for the upcoming Summer of Code project.

Talend is a open source [GPL] data integration tool used by companies to transform data from one format to another.

For example I might get 2-3 XML input files that I need to feed into a database, or SOLR server. It works really well until you start bumping into memory limits or file concerns when you handle large files.

Enter hadoop.

I'd would like to propose a project to write the necessary bats to make talend jobs run on a hadoop cluster, possibly using things like pig.

While I understand this code will probably end up as a part of talend's code base, I think it would be a neat project to expand hadoop's presence in this space.

I'm willing to act as a mentor for it. (I've been a mentor for HTP, and Lacre projects in the past).

regards
Ian
```

**Mime**

- Unnamed multipart/alternative (inline, None, 0 bytes)
  - Unnamed text/plain (inline, Base64, 1682 bytes)

View raw message
Our first Big Data project at Altic

- eFraudBox project (2010 – 2013)
  - Goal: predict frauds on Internet
  - Context:
    - Customer: GIE carte bancaire
    - European Research and Development project
    - Lot of industrial and academic partners
  - Data:
    - Type: Banking transactions
    - Volume: One GB per day
How did we start our first BigData project?
« In data mining processing is done line by line »

... [ there's not about a data volume issue ]
But we have too much data!
Let's have a look at Hadoop?

- Open Source
- MPP compute platform
  - Distributed file system
  - MapReduce processing
- Cost efficient
  - Fault tolerant
  - Infinite scale
- Enterprise Information System ready
- Continuous Improvement
- Growing community

« Even transactions are possible on Hadoop - it's inevitable that ALL kinds of workloads will move there in the future »

Doug CUTTING
Hadoop Creator
Octobre 2013
How do we query Hadoop?

- Java
- Very optimised
- Very customisable

- Pig Latin
- Easy syntax
- Support unstructured data

- SQL like
- Easy development
How do we query Hadoop?

- Need to code everything
- Why not?
- We already know SQL!
Ok, we have our storage and computation engine, but how can we manage data?

By using our **Swiss Army Knife**!
Now our Hadoop / Hive platform is filled with Big Data, but it's a little bit too slow to query for end users...
Aggregate data

Processing data with Hive and store results in fast databases

InfiniDB®
Ok, now we have our fast queryable datasets, but how can we visualize these?

- To manage users and visualizations
- To quickly have a vision of your data
- To go deeper in your visualizations

www.ow2.org  Twitter #ow2con @egwada
BigData and Datamining : tMahout

talend*
*open data solutions

= tMahout
BigData and Datamining v2

- Spark: new InMemory data processing framework
  - Very appropriate for Machine learning
  - MLBase: Machine learning library
  - Spark-clustering: Implementation of SOM algorithm
  - Proof Of Concept: Analysis of mobile telecommunications
We have now a Big Data stack!
BI & Big Data for Altic

- Eventually, we still do BI as usual
  - Tools evolve:
    - New storage and processing
    - We do not change our tools, fortunately THEY progress for us and we contribute
  - Fundamental does not really change, only technologies do
    - Hadoop
    - Spark
We improve our Big Data stack and its approach...

And support Big Analytic customer project

Our Big Data Stack

Our Big Data Approach
Questions?

Thanks!

Charly CLAIRMONT
CTO at ALTIC

@egwada
charly.clairmont@altic.org

http://altic.org