REX Hadoop Air France

01/07/2014

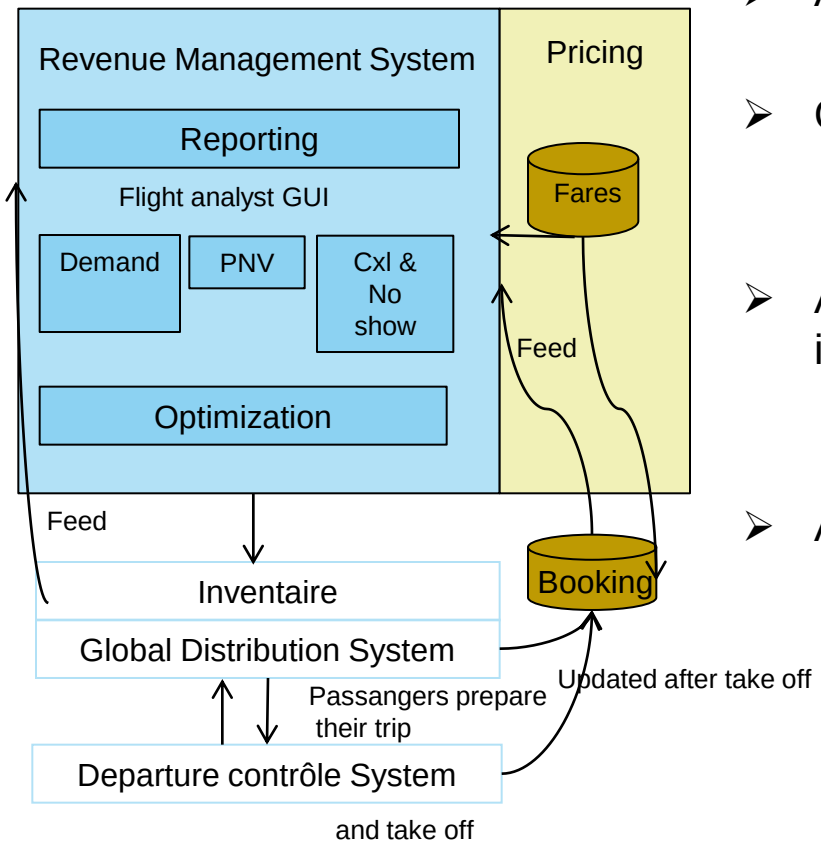
J.Maréchal

01 Agenda

- Use case context.
- Infrastructure approach.
- Hadoop.
- Feedback.

AIRFRANCE KLM

Application Use Case Context



- AF/KLM RMS (Revenue Management System).
- Optimization based on :
  - Demand, Cancellation, Overbooking
- Application can also let the flight analyst to interact on recommendation based on :
  - Markets, Periods, Events ...
- Application run on 3 domains :
  - User activity
  - Batch activity (nightly batch)
  - Event activity

Application infrastructure

- Standards technical frameworks : Linux, BD, AS, HTTP, NFS, Java Application and Operational research, C++/Cplex.

- New technical frameworks : Hadoop and Grid Engine.

  - Grid Engine (OGS) : grid management for batch launched by transactionnal part.
  - Hadoop : bacth activity by delivering an important parallelism.

Infrastructure requirements

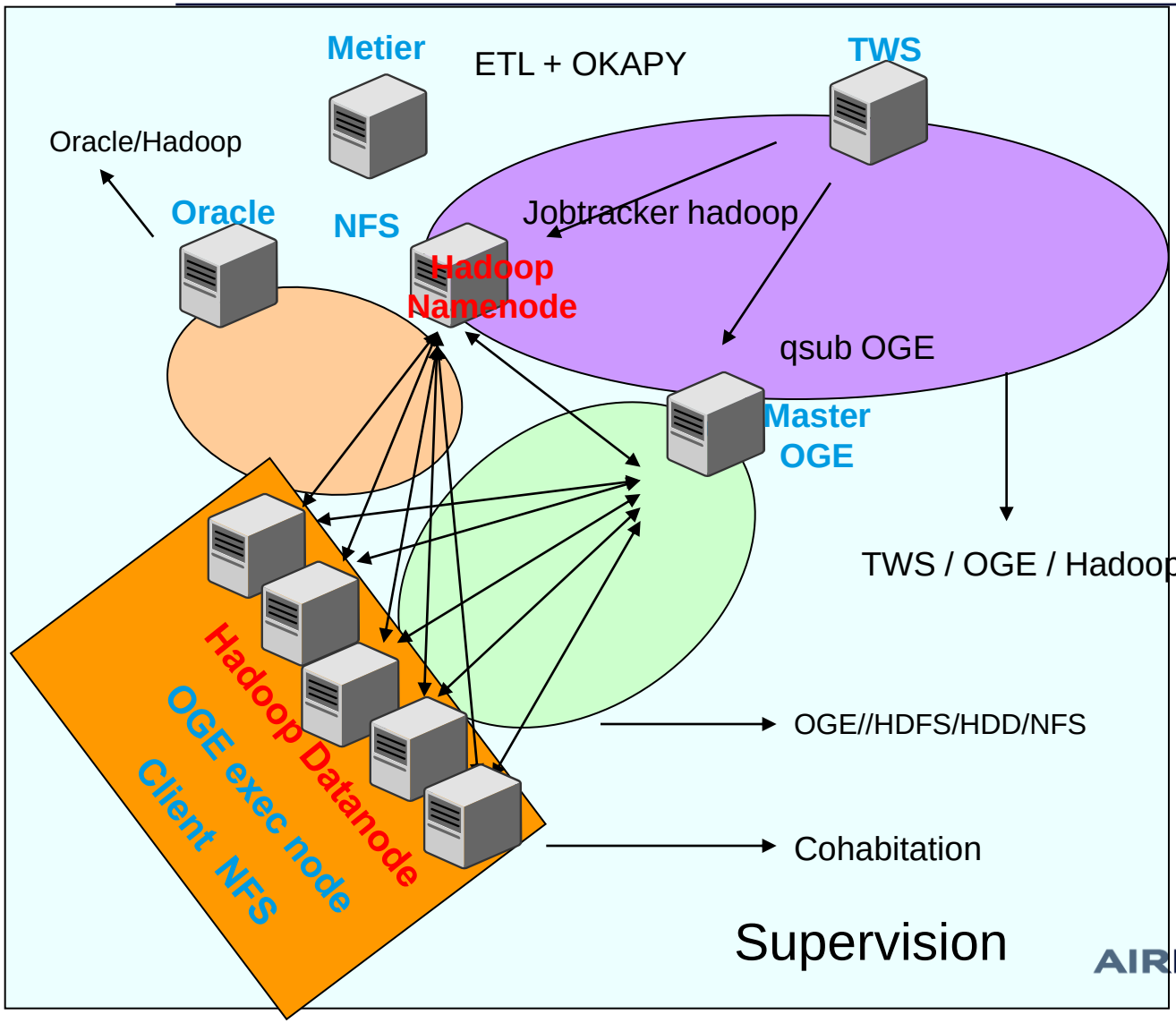
- Hadoop integrated and managed as a technical framework.

- OPS are responsible of :

  - The SLA of the framework in the active/active DRP context.
  - Operated 24*7 by the OPS (level 1 and level 2).
  - Monitored, RPO=0
  - RTO based on the application SLA.
  - The integration on the standards stack
  - The Life Cycle Management.

➔ Applications using it are considered as a customer.

Schema infrastructure technique



Infrastructure

- 2 sockets R720XD
- E5-2670 8 cores
- Mem 128GB

- 21 nodes per envt

- 75To hadoop JBOD storage per envt.

- 5 envts

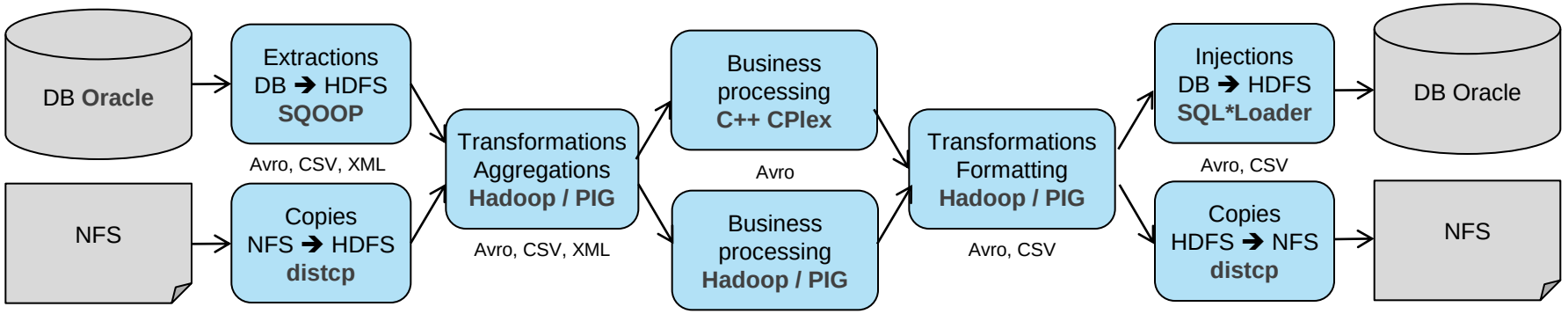
Hadoop OPS facts in 2011

➢ No unified framework management.	⟶	❑ Homemade Admin framework
➢ No backup tool.	⟶	❑ Hdfs to NFS and reload
➢ No HA functionnalities	⟶	❑ Symantec for name node.
➢ No monitoring tool.	⟶	❑ HPOV
➢ No performance reporting.	⟶	❑ Ganglia
➢ No hadoop consolidation available	⟶	❑ Multi-context
➢ No automated developpement	⟶	❑ Deployment automated through the internal cloud.

AIRFRANCE KLM

Hadoop integration.

- Used for the nightly batch activity (high volumetry, and performance contraints).
- Module used :
  - Map/reduce, HDFS, Pig, Avro, Sqoop.
  - HDFS only use has a support of the datas for performance and parallélism.
  - Simplify application coding by structuring the datas post threatment.



Feedback 1/2

- Tunning : cpu bound, I/O bound or memory bound ?
- Data to HDFS
  - Used of compressions codecs.
  - Compactor to optimize the volume and rise performance.

- BD and HDFS :
  - Sqoop BD to HDFS
  - SQL*loader HDFS→ BD

- Tool developed to research easily the datas through a GUI.
- Pig usage for KPI and validate the threatments.

Feedback 2/2

- Difficulty to do the link between treatments, map/reduce task and log.
- Difficulties of configuration between treatment and bundle with the environment (slot number, RAM, specific option).
- Difficulties to define the best number of map slots and reduce slots.
- Manage defragmentation of the HDFS.

Conclusion

- Define your Big Data approach (opensource mainstream, appliance, opensource edition)
  ➔Use the enterprise edition
- Define how you want to organize the data in the HDFS.
- Need specialized ressources
  ➔Need strong collaboration (OPS & DEV).

AIRFRANCE KLM