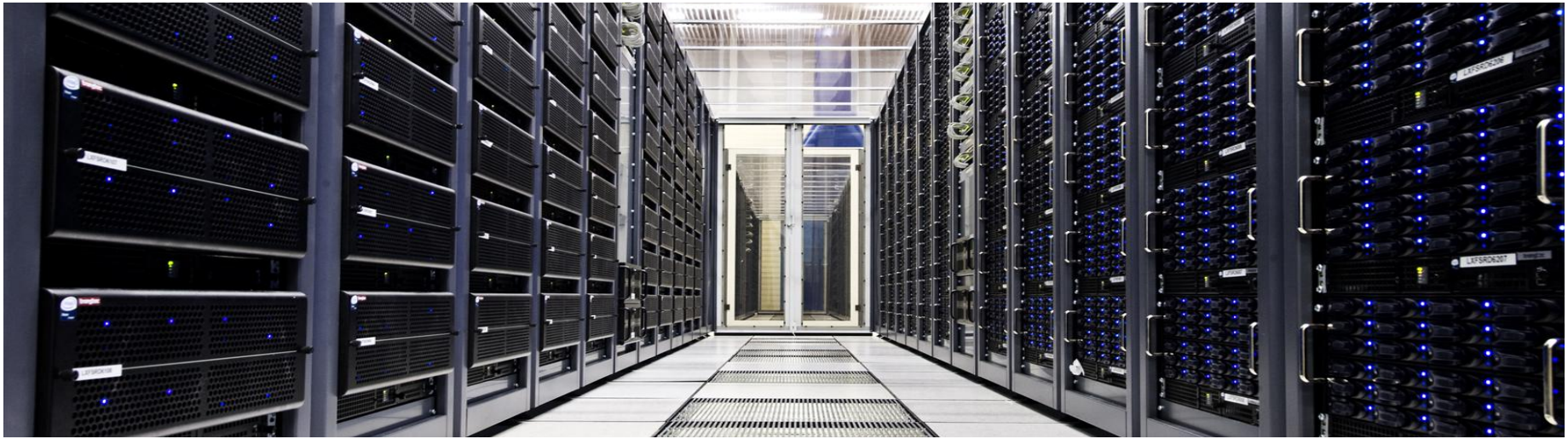
Bridging Oracle with Hadoop

Zbigniew.Baranowski@CERN.CH

XLDB, SLAC, May 2016



Why integrate Oracle with Hadoop?

- Oracle: Online Transactional System
- Hadoop: Large Scalable Data Warehouse

- Build a hybrid systems:
  - Move (read-only) data from Oracle to Hadoop
  - Query Hadoop data from Oracle

- Increase scalability and lower ratio cost/performance
  - Hadoop data formats and engines for high performance analytics
  - ….without need of changing the end-user apps connecting to Oracle

Example

```
select
    date,
    count(*)
from big_table @ oracle
    where…
    group by date
```

Query will run for about 3 days (example based on 2GB/s)

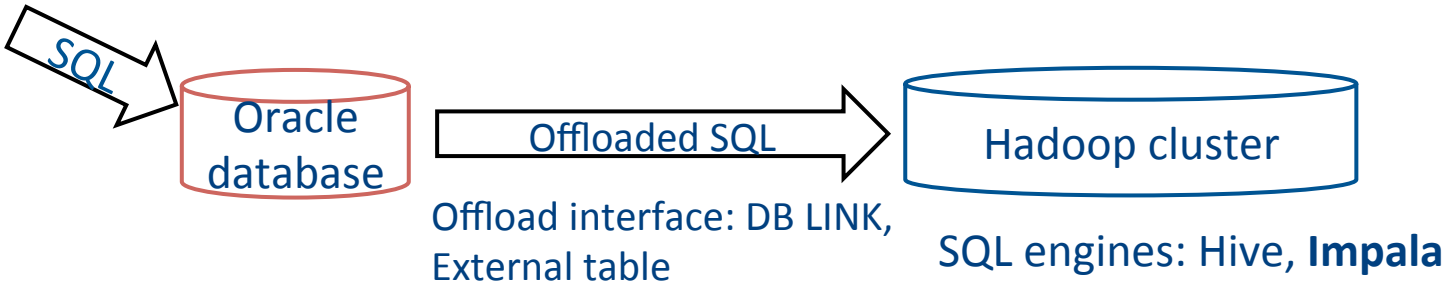
500TB of data

Various filter predicates

Integrating Oracle and Hadoop

- Step1: Offload data to Hadoop

- Step2: Offload queries to Hadoop (full or partial)

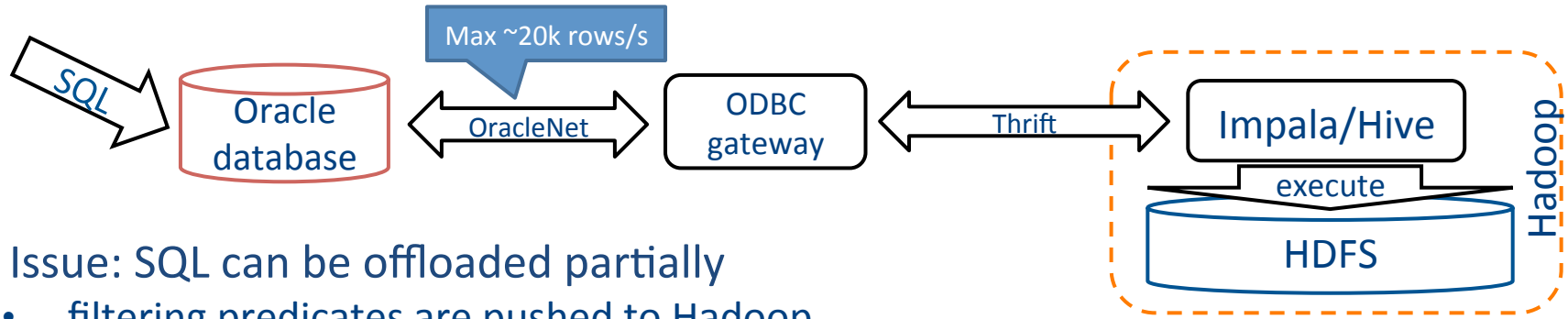


Solution with no additional cost

- Query Apache Hive/Impala tables using a database link

```
create database link my_hadoop using 'impala-gateway';
select count(*) from big_table@my_hadoop;
```

- Query offloaded via ODBC gateway to Impala (or Hive)



- Issue: SQL can be offloaded partially
  - filtering predicates are pushed to Hadoop
  - grouping aggregates are not pushed!
- There are techniques to workaround this problem
  - create aggregation views in Hive/Impala
  - DBMS_HS_PASSTHROUGH – to push exact SQL statement to Hadoop

Making data sources transparent to end-user

- Hybrid views on Oracle
  - recent (read-write) data in Oracle
  - archive data in Hadoop

Split point has to be updated after each successful data offload

```
create view big_table as
    select * from oracle_online_table where date > '2016-05-25'
    union all
    select * from archival_big_table@hadoop where date <= '2016-05-25'
```

Specialized products for hybrid and offloads

- Oracle SQL connectors for Hadoop
  - no Hadoop-side processing

- Oracle BigData SQL
  - now available for non-BDA installations
  - custom engine for Hadoop-side data filtering

- Gluent Inc
  - uses Apache Impala to process data on Hadoop
  - leverage hybrid views on Oracle for data integrity
  - implicit predicates pushing and partition pruning
  - data retrieval >10x faster than ODBC gateway

Summary

- Solutions available for querying Hadoop from Oracle
  - pushing SQL
  - hybrid views on top of partitioned tables (Oracle + Hadoop)
  - remote access using DB link or external table
  - no silver bullet

- Hybrid systems (Oracle + Hadoop) to
  - lower cost/performance ratio for analytic workloads
  - profit from Oracle for OLTP and Hadoop for DW