Impala

Impala: Goals

- General-purpose SQL query engine for Hadoop
- High performance
  - C++ implementation
  - runtime code generation (using LLVM)
  - direct data access (no MapReduce jobs)
- Run directly on Hadoop
  - read the same file formats
  - use the same storage managers (Hive metastore)
  - daemons on the same nodes that run Hadoop processes

Data formats

- Supported HDFS file formats
  - Parquet
  - Text
  - Avro*
  - RCFile*
  - SequenceFile*

  * no inserts, use Hive for that
- Querying HBase tables possible
- Querying Amazon S3 Filesystem in test phase

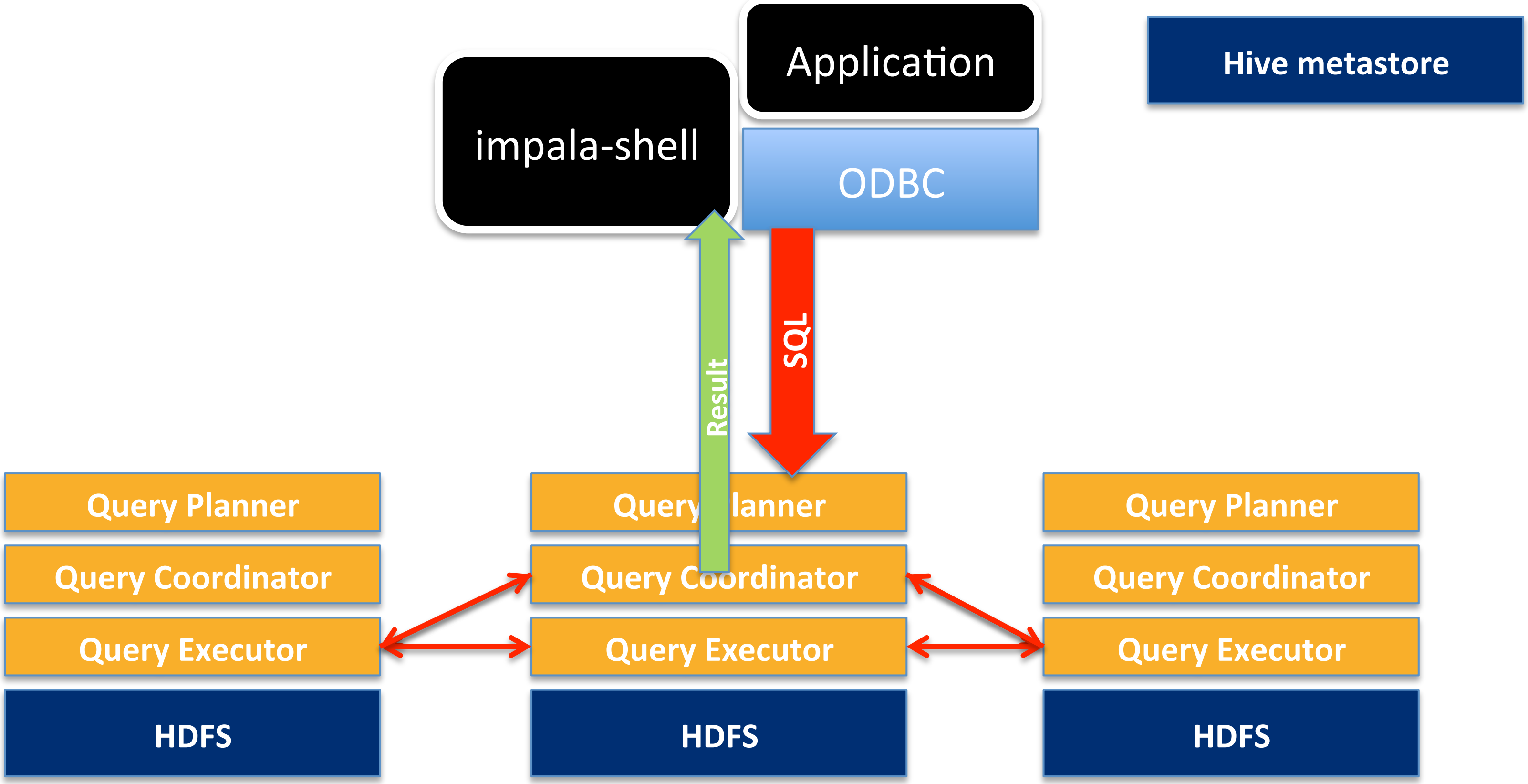
User interfaces

- impala-shell for interactive commands
- Apache Hue as web-based user interface
- JDBC and ODBC to connect from applications
  - or as external database from Oracle

Components

- impala daemon (`impalad`)
  - one pre node
  - accepts queries, distributes work, transfers results back to the coordinator node
- impala statestore (`statestored`)
  - one pre cluster
  - monitors health of impala daemons
- impala catalog service (`catalogd`)
  - one pre cluster
  - transfers metadata changes from impala sql statements

Query execution



Impala metadata and Hive metastore

- table definitions in shared Hive metastore
- impala tracks additional metadata inc.:
  - physical location of blocks in HDFS
- after external changes (through Hive or manually to files) metadata needs to be updated
  - REFRESH table_name, INVALIDATE_METADATA

Hands on: Impala & Hive metastore

1. Create table in Impala
   - check if it’s accessible in Hive
   - check content of default Hive folder
   - try inserting
2. Vice versa. Create table in Hive
   - check if it’s accessible in Impala
   - try inserting

commands: http://cern.ch/kacper/impala1.txt

Query optimizer

- Commands available for performance tuning
  - `EXPLAIN SELECT`… - steps that a query will perform
  - `SUMMARY` – report about the last executed query
  - `PROFILE` – like `SUMMARY` but more detailed and low-level information
- Table statistics are stored in Metastore
  - can be viewed using
    - `SHOW TABLE STATS table_name`
    - `SHOW COLUMN STATS table_name`
  - if missing, use
    - `COMPUTE STATS table_name`

Hands on: Table stats and Explain

1. Prepare weather calculation query
2. Check current table statistics
3. View execution plan
4. Compute statistics
   - check for differences in statistics
   - check for differences in the execution plan
5. Have a look on summary and profile

commands: http://cern.ch/kacper/impala2.txt