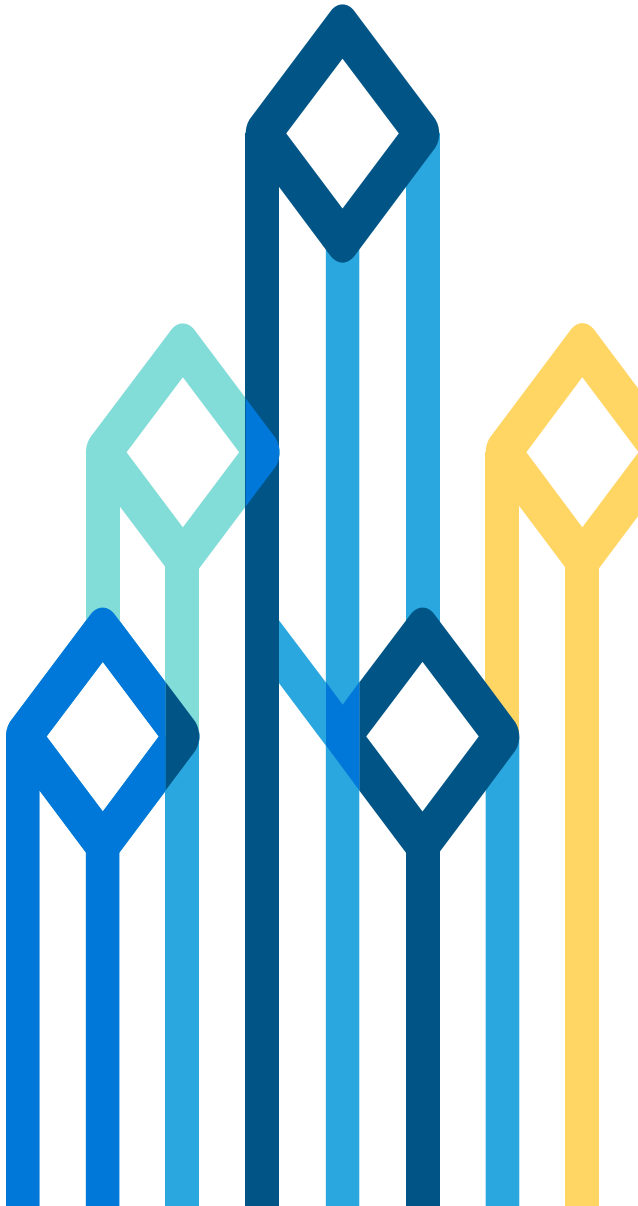
cloudera

Big Data Analytics with Tableau and Apache Impala

The Leader for Analytic SQL on Hadoop

Scott Armstrong, Director, Cloudera
Murali Krishna, CTO Colaberry
Madhu Ganta, Big Data Architect, Cloudera
Jatin Shah, Sr. Sales Engineer, Cloudera

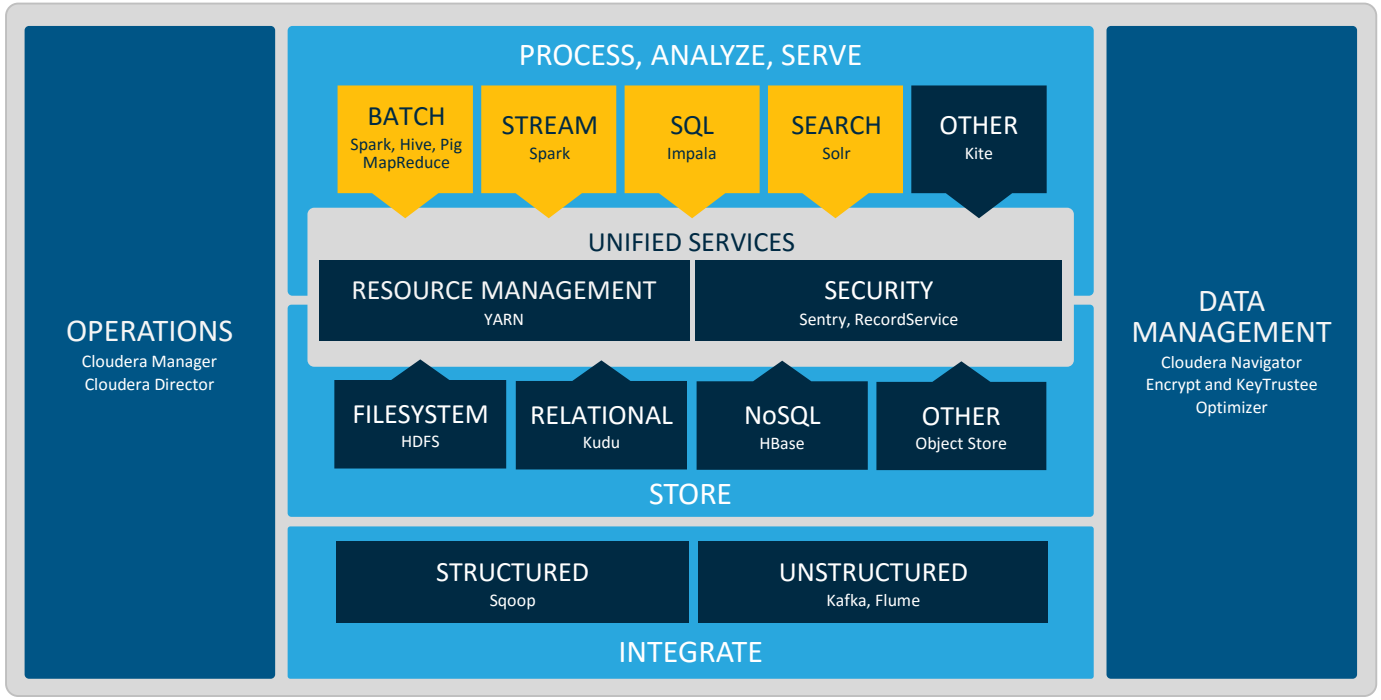


Solution: Hadoop & Ecosystem

- Hadoop is a software platform for storing, processing, and analyzing “big data”
  - Distributed
  - Scalable
  - Fault-tolerant
  - Open source

© Cloudera, Inc. All rights reserved.

One Platform, Many Workloads



Batch, Interactive, and Real-Time.

- End-to-end analytic workflows
- Access more data
- Work with data in new ways
- Enable new users

© Cloudera, Inc. All rights reserved.

One Platform. Many Applications.

© Cloudera, Inc. All rights reserved.

Apache Impala

© Cloudera, Inc. All rights reserved.

What is Impala ?

- Apache Impala is a high-performance, low-latency SQL queries engine on data stored on the popular Apache Hadoop file system (HDFS) .

- Impala integrates with the existing CDH ecosystem & shares same hive meta-store

© Cloudera, Inc. All rights reserved.

Apache Impala: Open Source & Open Standard

1. > 1 MM downloads since GA
2. Majority adoption across Cloudera customers
3. Certification across key application partners:

4. De facto standard with multi-vendor support:

© Cloudera, Inc. All rights reserved.

Database Cost per TB

MPP Database

- Massively Parallel Processing
- Specialty Database Appliances
- $30-$50K per TB
- Expensive

Apache Impala on Hadoop

- Massively Parallel Processing
- Really all Commodity Hardware
- $2K per TB
- At least 10x Cheaper

© Cloudera, Inc. All rights reserved.

What is it good at?

© Cloudera, Inc. All rights reserved.

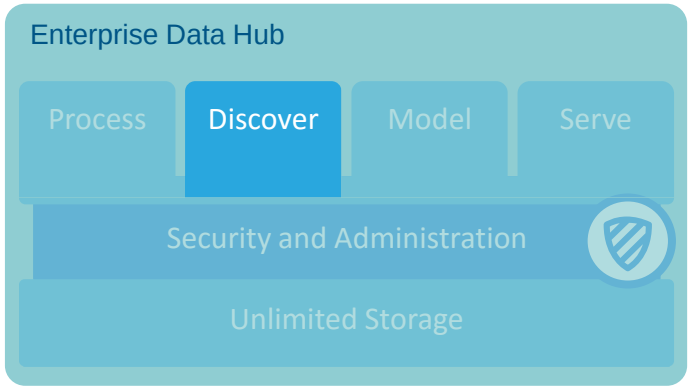
Good at

- Schema on read – changing schemas
- Columnar Compression
- Ugly Wide tables
- Big data that doesn’t fit anywhere else
- Interactive analysis at Scale
- Performance at 5-10% of the cost
- Data Consolidation – Data Lake, Data Hub

© Cloudera, Inc. All rights reserved.

Successful Use Cases

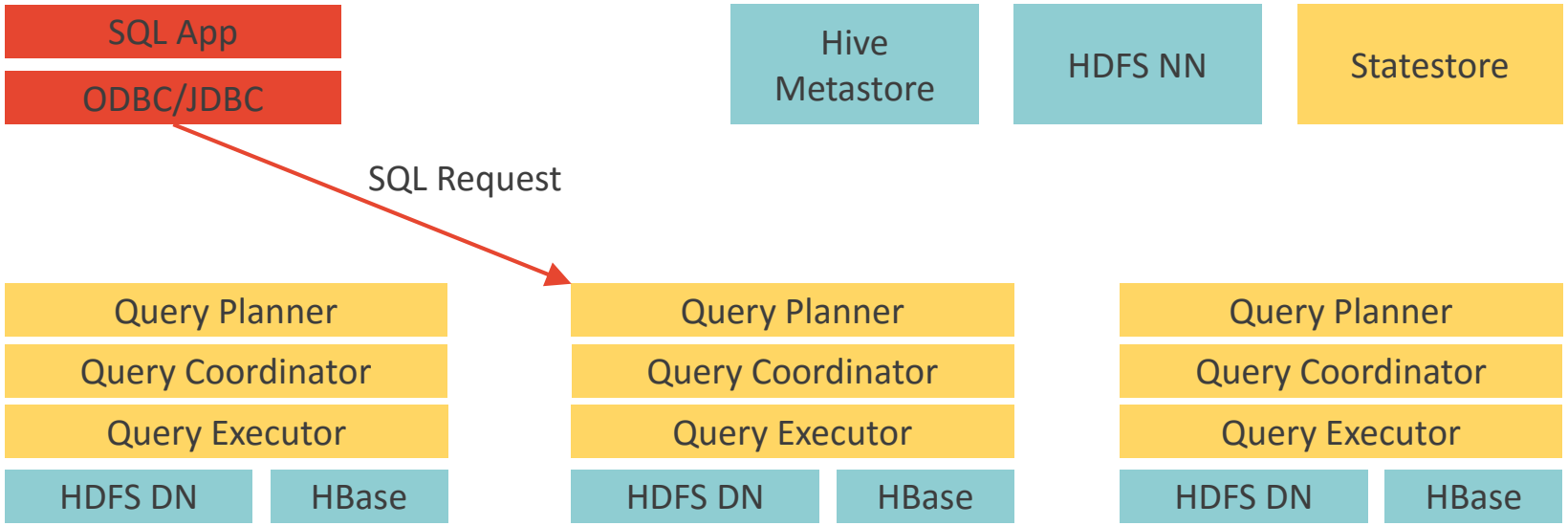
- Interactive BI/analytics on “big data”
- Data discovery – Understanding your data
- Exploratory analytics – Top N , Percentiles etc.
- Queryable operational data store
- Active Archive



© Cloudera, Inc. All rights reserved.

Impala Architecture

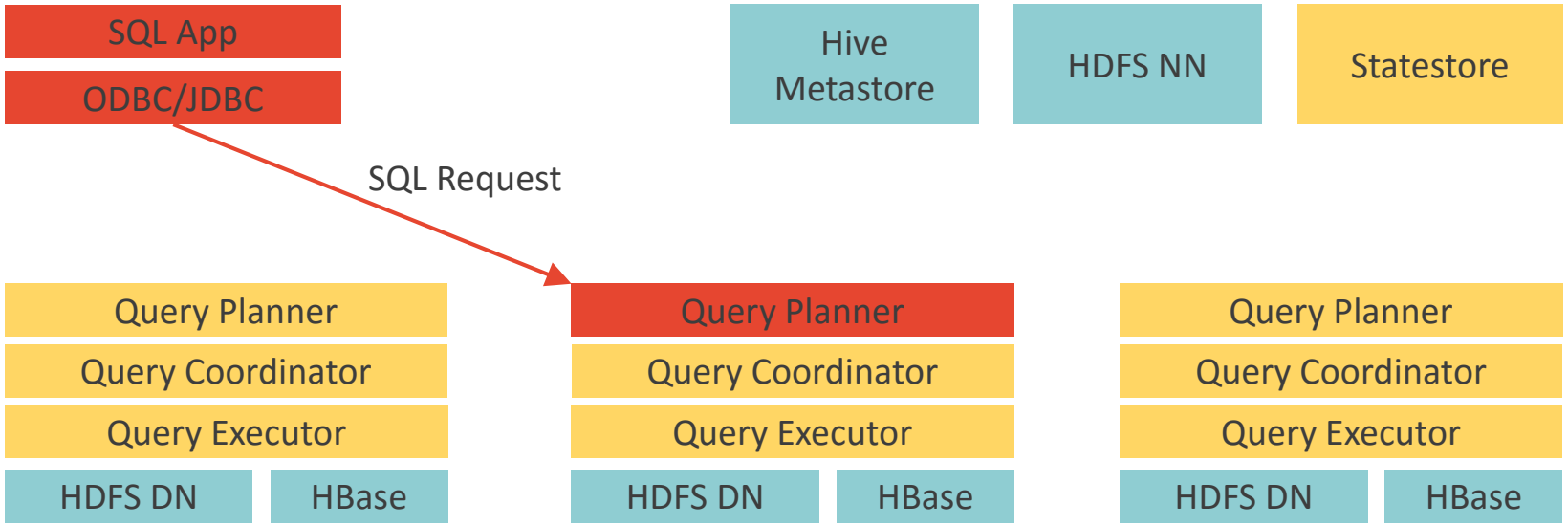
- MPP query engine built natively into Hadoop



© Cloudera, Inc. All rights reserved.

Impala Architecture: Query Execution

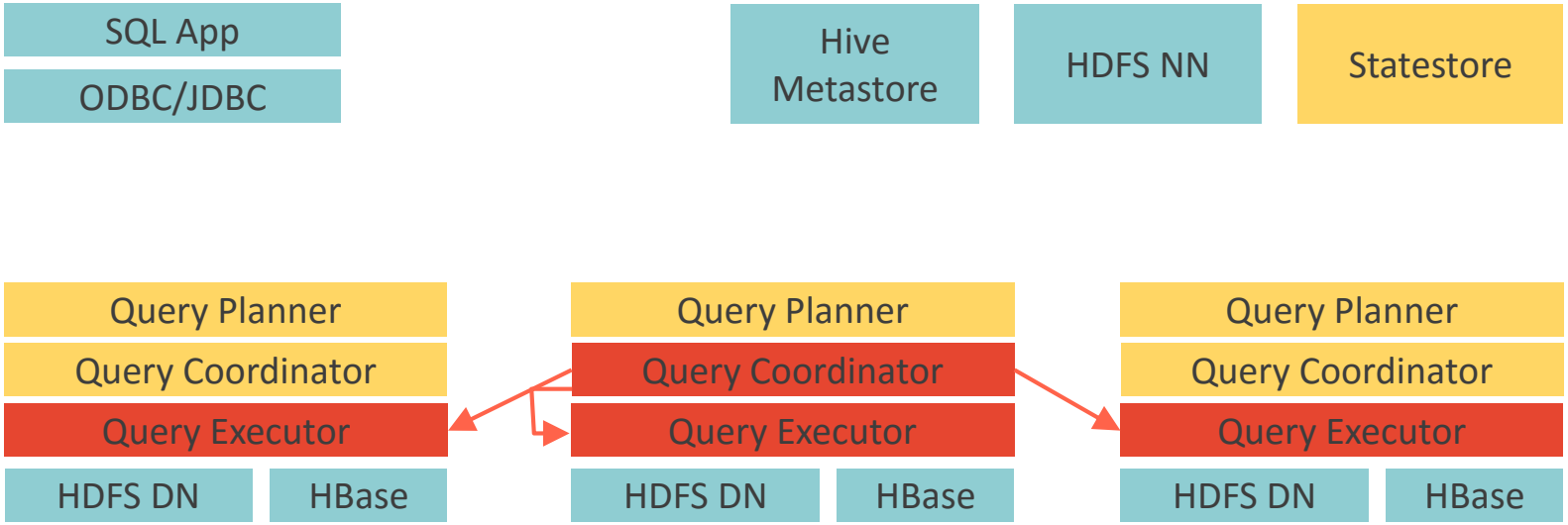
- Request arrives via ODBC/JDBC/Hue GUI/Shell



© Cloudera, Inc. All rights reserved.

Impala Architecture: Query Execution

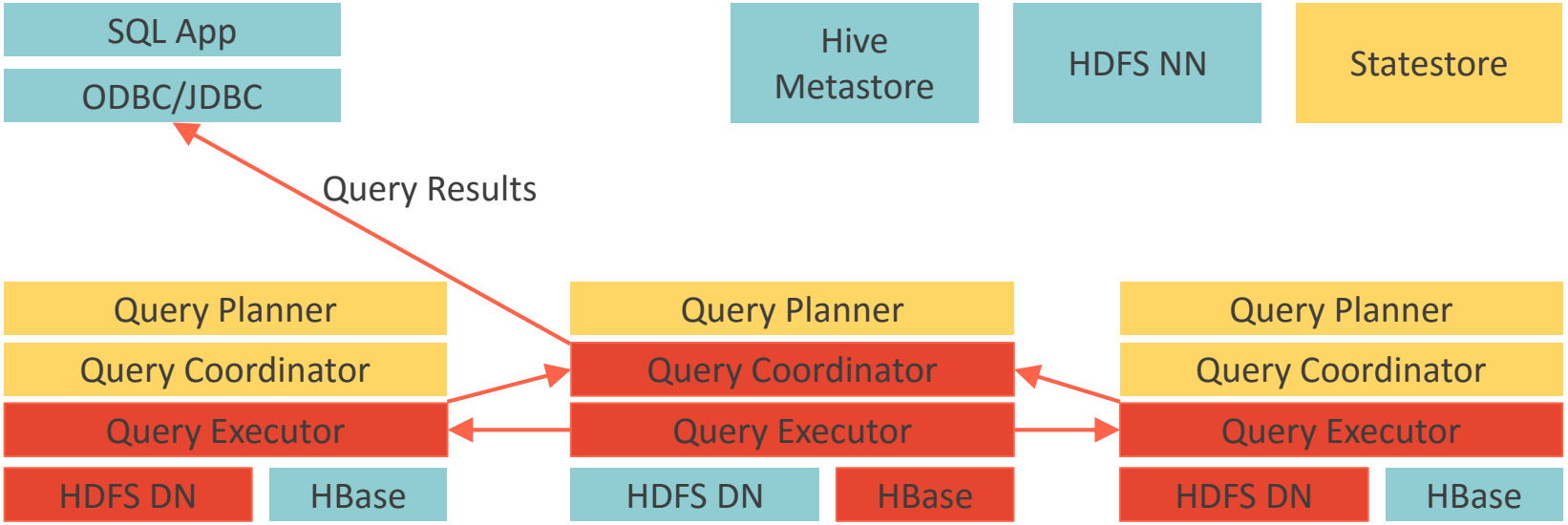
- Planner turns request into collections of plan fragments
- Coordinator initiates execution on impalad's local to data



© Cloudera, Inc. All rights reserved.

Impala Architecture: Query Execution

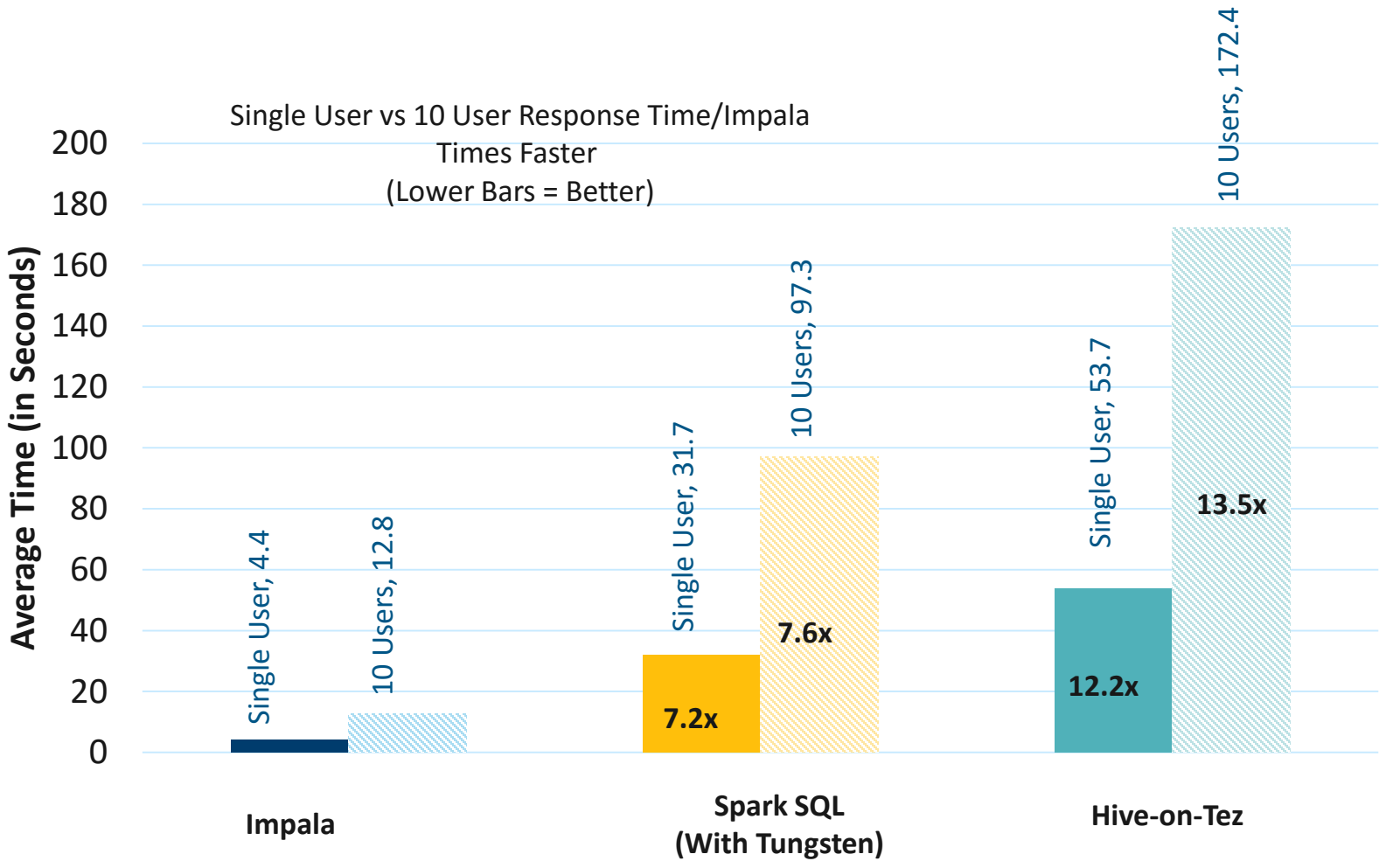
- Intermediate results are streamed between impalad's
- Query results are streamed back to client



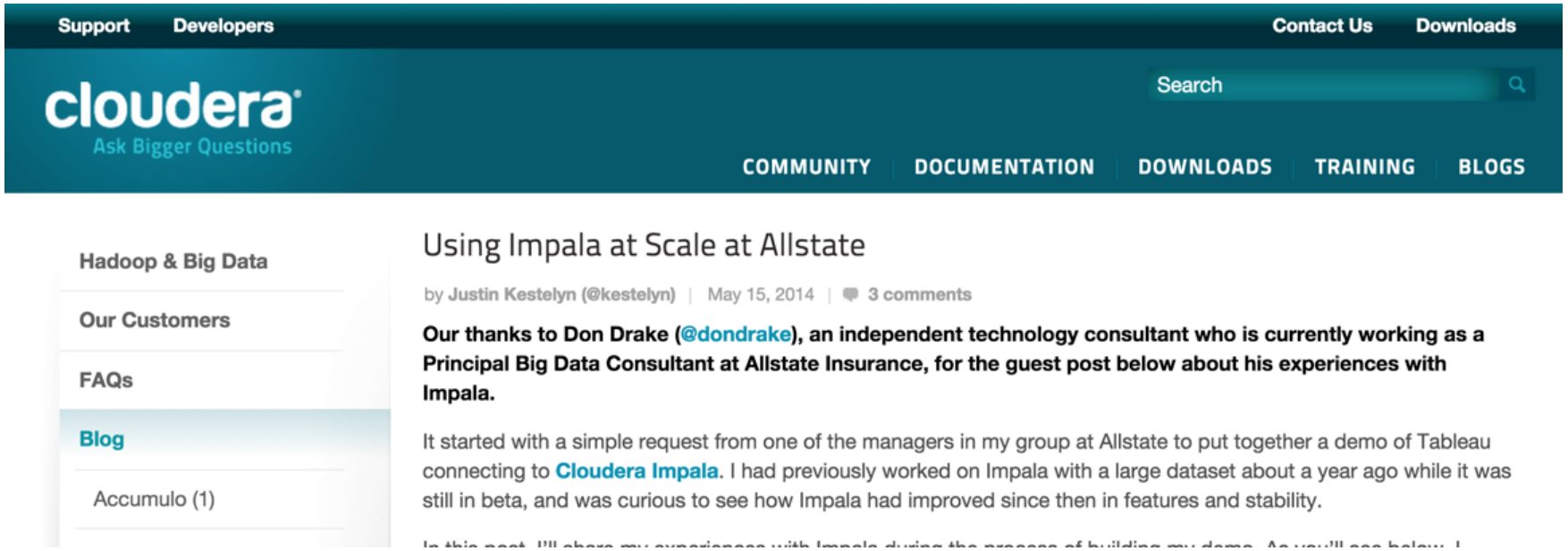
© Cloudera, Inc. All rights reserved.

Performance

© Cloudera, Inc. All rights reserved.



© Cloudera, Inc. All rights reserved.



- http://blog.cloudera.com/blog/2014/05/using-impala-at-scale-at-allstate/

© Cloudera, Inc. All rights reserved.

Blog Summary

- Earned Premium Dataset – 800 columns, 1.1B Rows, 2.3TB
- 35 node cluster – scans text table in 3 minutes
- Add Partitioning (time, geo, org, line) and Parquet Format
- Runs like lightning – where'd my 2TB go? Now 100GB
- Can scan the whole table in 10 seconds
- Partitioned queries run in 1 or 2 seconds

```
DROP TABLE IF EXISTS default.datasetp_2010_2013;
CREATE EXTERNAL TABLE IF NOT EXISTS default.datasetp_2010_2013
(
PROC_MONTH STRING, PROC_YEAR STRING, POLCT STRING, STATST STRING, [800+ column definit
EEXP REAL, EPREM REAL, EXP REAL, PREM REAL
)
PARTITIONED BY (ACTYR STRING, GEOST STRING, ALINE STRING, COMPNY STRING)
STORED AS PARQUET
LOCATION '/user/drake/cdf_impala_part';
```


© Cloudera, Inc. All rights reserved.

Tuning & Best Practices

© Cloudera, Inc. All rights reserved.

Partitioning

- Large tables can be partitioned
- Common partition field is date
- Partitions should be > 1GB
- Common where clause criteria

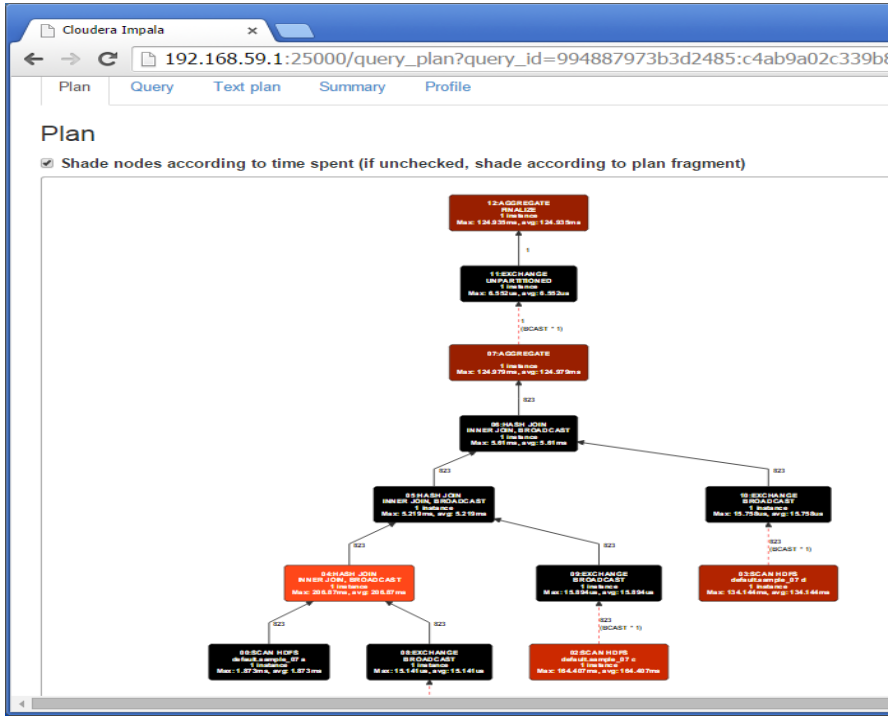
© Cloudera, Inc. All rights reserved.

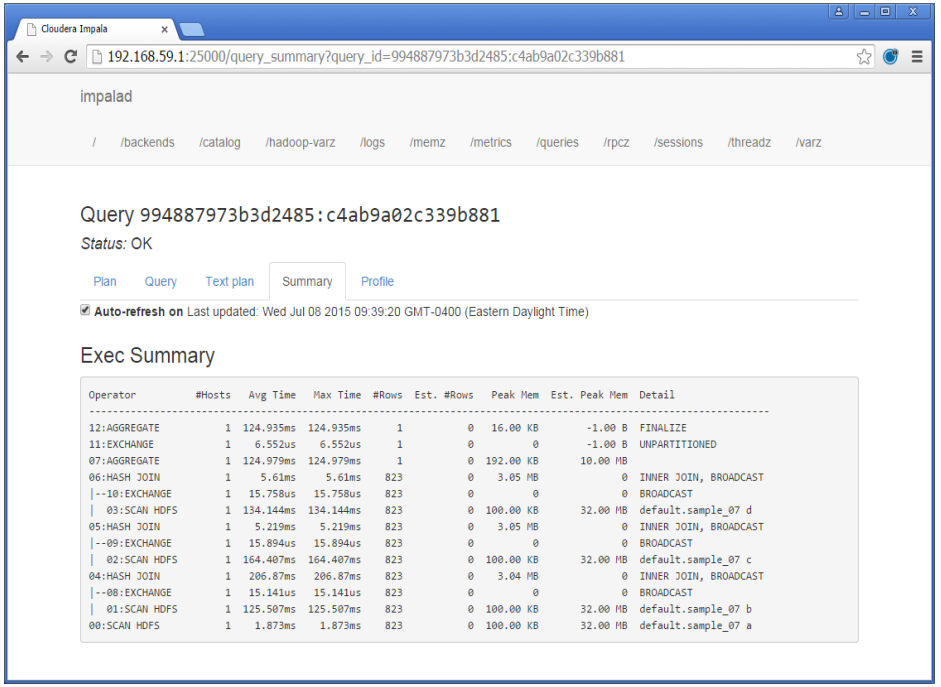
Why create a bunch of tables?

- 80% of data modeling is to save space or create performance
- Making Impala do a join is often a waste of time
- Parquet lets you build big ugly tables that work well
- Lazy data modeling advocate

© Cloudera, Inc. All rights reserved.

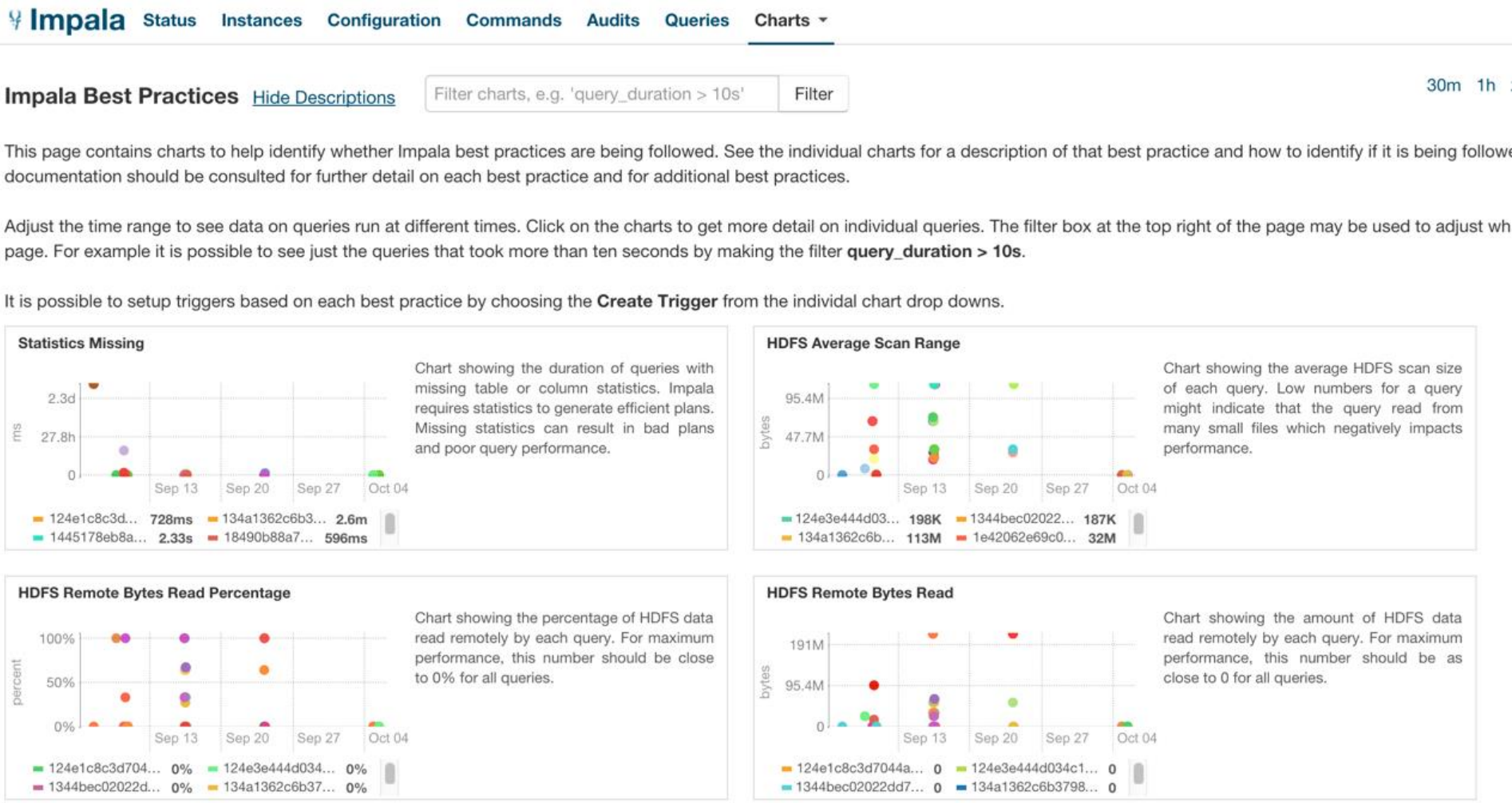
Impala query analysis





Operator	#Hosts	Avg Time	Max Time	#Rows	Est. #Rows	Peak Mem	Est. Peak Mem	Detail
12:AGGREGATE	1	124.935ms	124.935ms	1	0	16.00 KB	-1.00 B	FINALIZE
11:EXCHANGE	1	6.552us	6.552us	1	0	0	-1.00 B	UNPARTITIONED
07:AGGREGATE	1	124.979ms	124.979ms	1	0	192.00 KB	10.00 MB	
06:HASH JOIN	1	5.61ms	5.61ms	823	0	3.05 MB	0	INNER JOIN, BROADCAST
|--10:EXCHANGE	1	15.758us	15.758us	823	0	0	0	BROADCAST
| 03:SCAN HDFS	1	134.144ms	134.144ms	823	0	100.00 KB	32.00 MB	default.sample_07 d
05:HASH JOIN	1	5.219ms	5.219ms	823	0	3.05 MB	0	INNER JOIN, BROADCAST
|--09:EXCHANGE	1	15.894us	15.894us	823	0	0	0	BROADCAST
| 02:SCAN HDFS	1	164.407ms	164.407ms	823	0	100.00 KB	32.00 MB	default.sample_07 c
04:HASH JOIN	1	206.87ms	206.87ms	823	0	3.04 MB	0	INNER JOIN, BROADCAST
|--08:EXCHANGE	1	15.141us	15.141us	823	0	0	0	BROADCAST
| 01:SCAN HDFS	1	125.507ms	125.507ms	823	0	100.00 KB	32.00 MB	default.sample_07 b
00:SCAN HDFS	1	1.873ms	1.873ms	823	0	100.00 KB	32.00 MB	default.sample_07 a

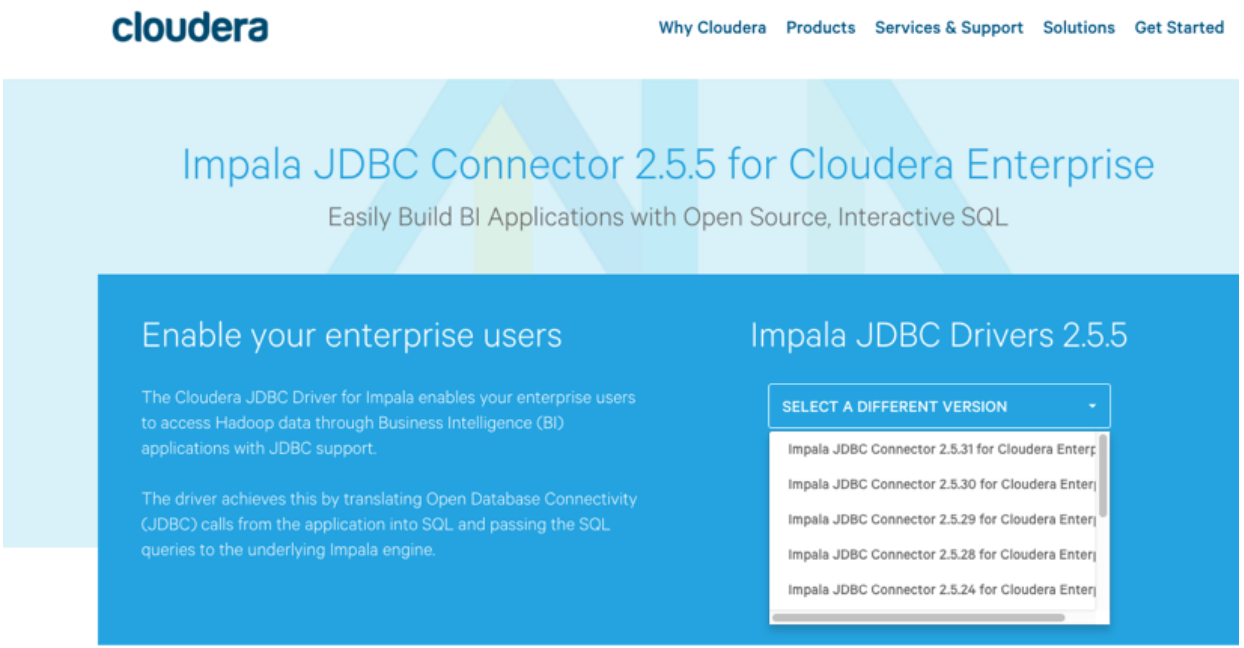
© Cloudera, Inc. All rights reserved.



© Cloudera, Inc. All rights reserved.

Cloudera ODBC and JDBC Drivers

- Written by Simba Technologies
- Standard BI data access



© Cloudera, Inc. All rights reserved.

Tableau

URL: ec2-54-201-244-144.us-west-2.compute.amazonaws.com
Port: 21050

© Cloudera, Inc. All rights reserved.

Demo

© Cloudera, Inc. All rights reserved.

cloudera®

Thank You

jshah@Cloudera.com

madhu@cloudera.com

murali@colaberry.com

© Cloudera, Inc. All rights reserved.