

# Hadoop Data Warehouse for BI & Analytics

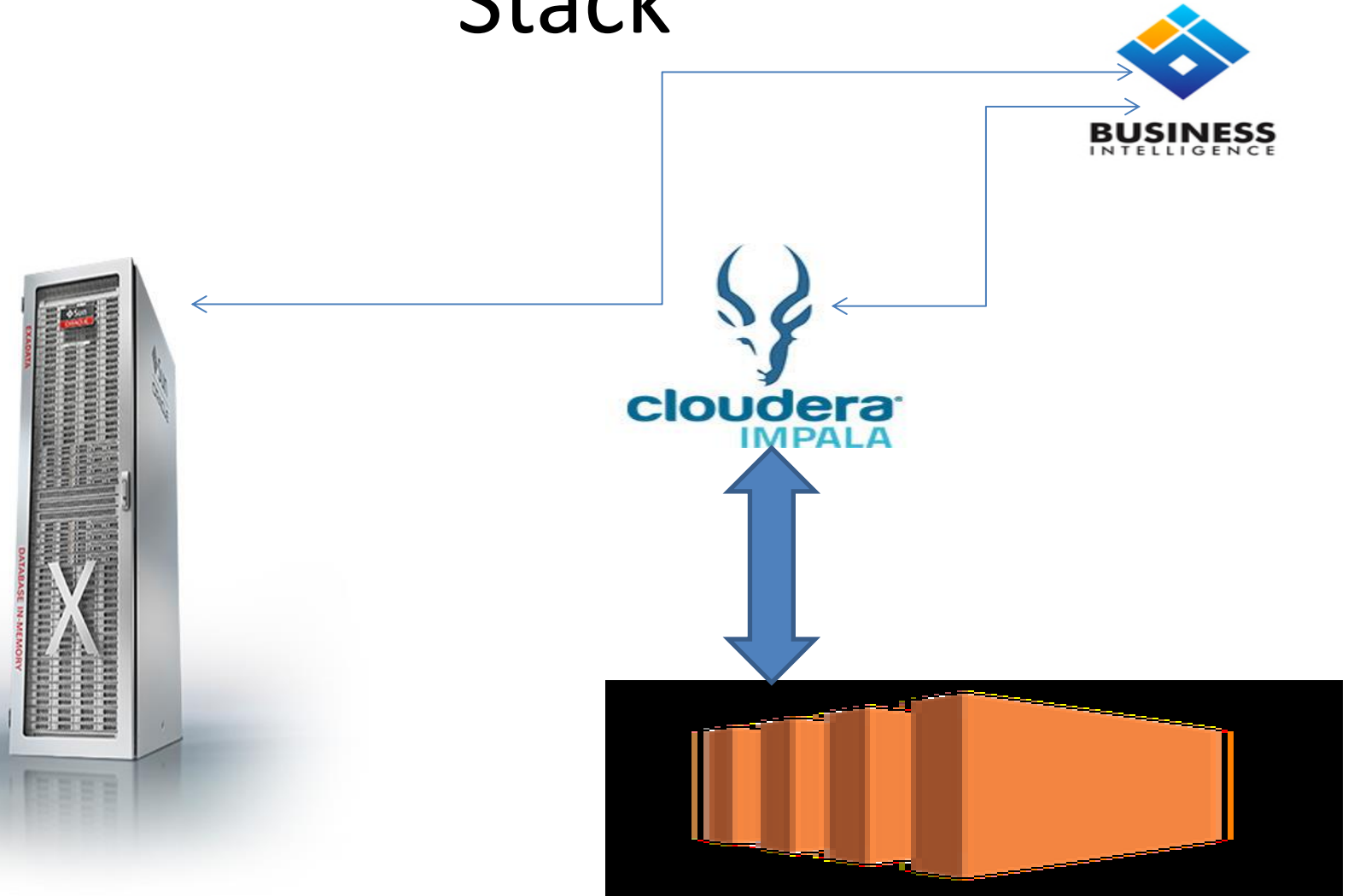
Kalyana Miriyala



# Motivation

- Build a scalable and performant Data warehouse on Hadoop for BI & Analytics
  - accommodate slowly changing dimensions

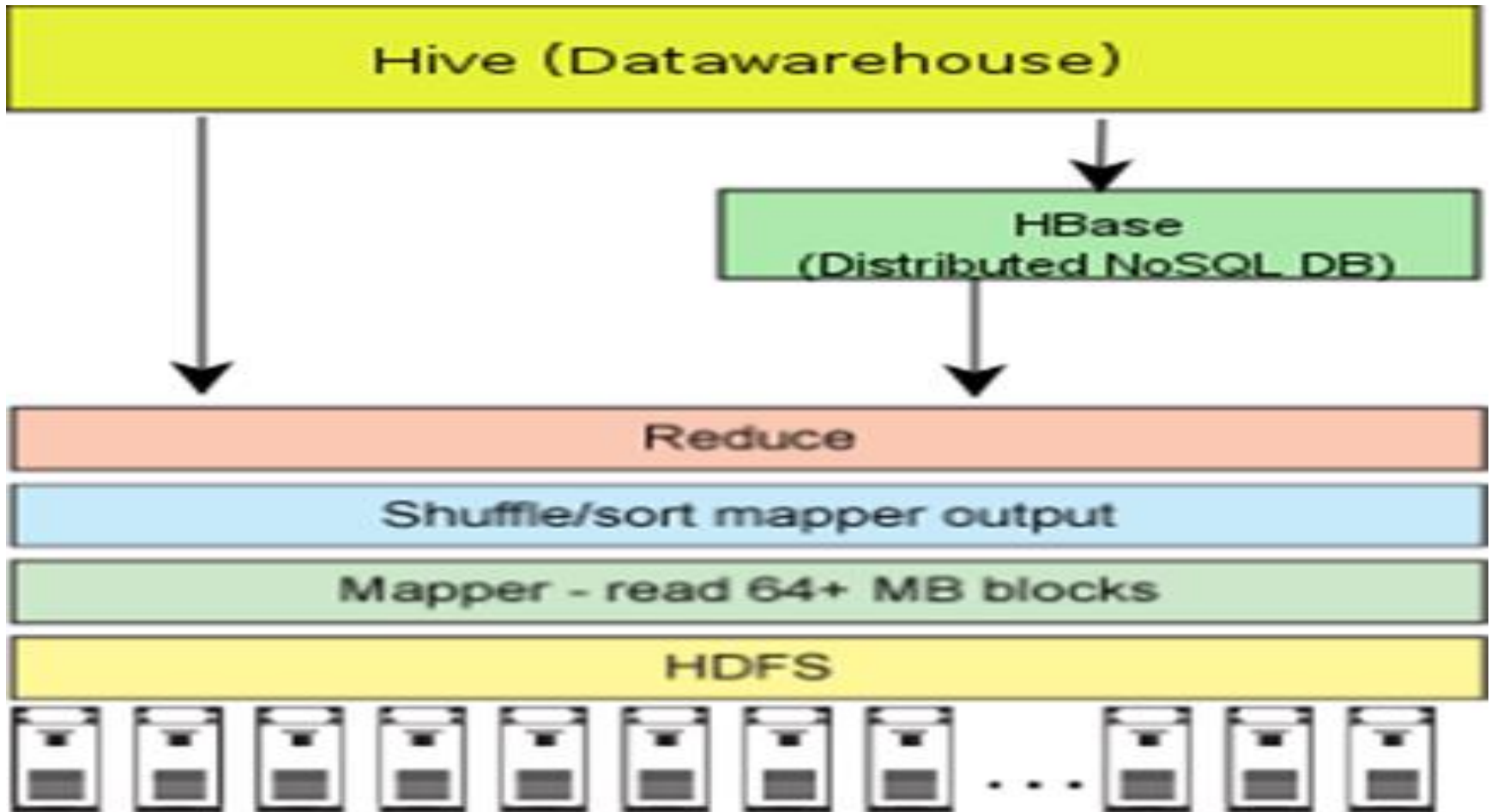
# Stack



Oracle Exadata Appliance

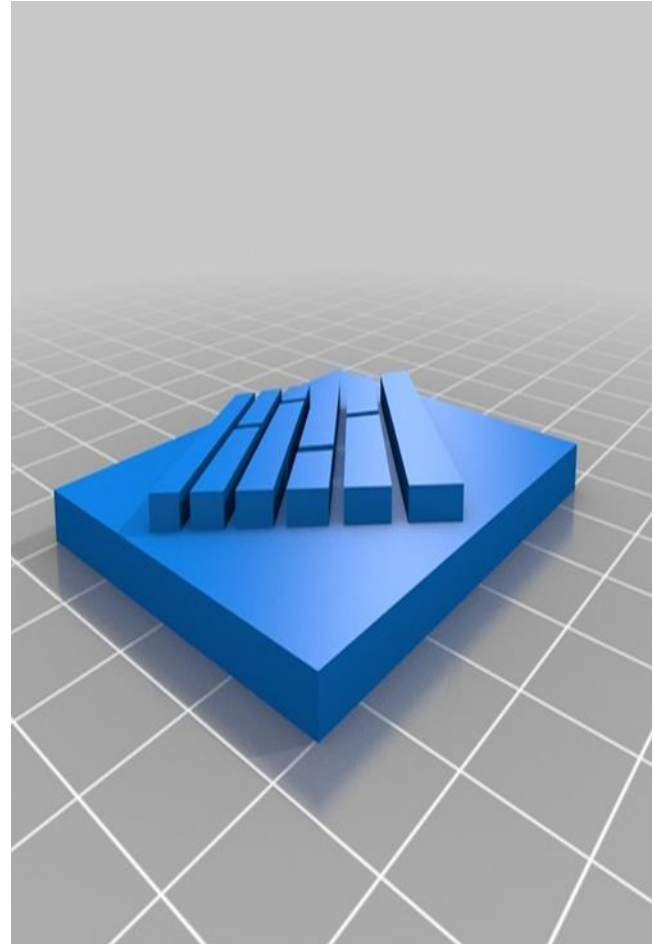
AWS r3 8Xlarge 4 Node Cluster

# Data tier stack



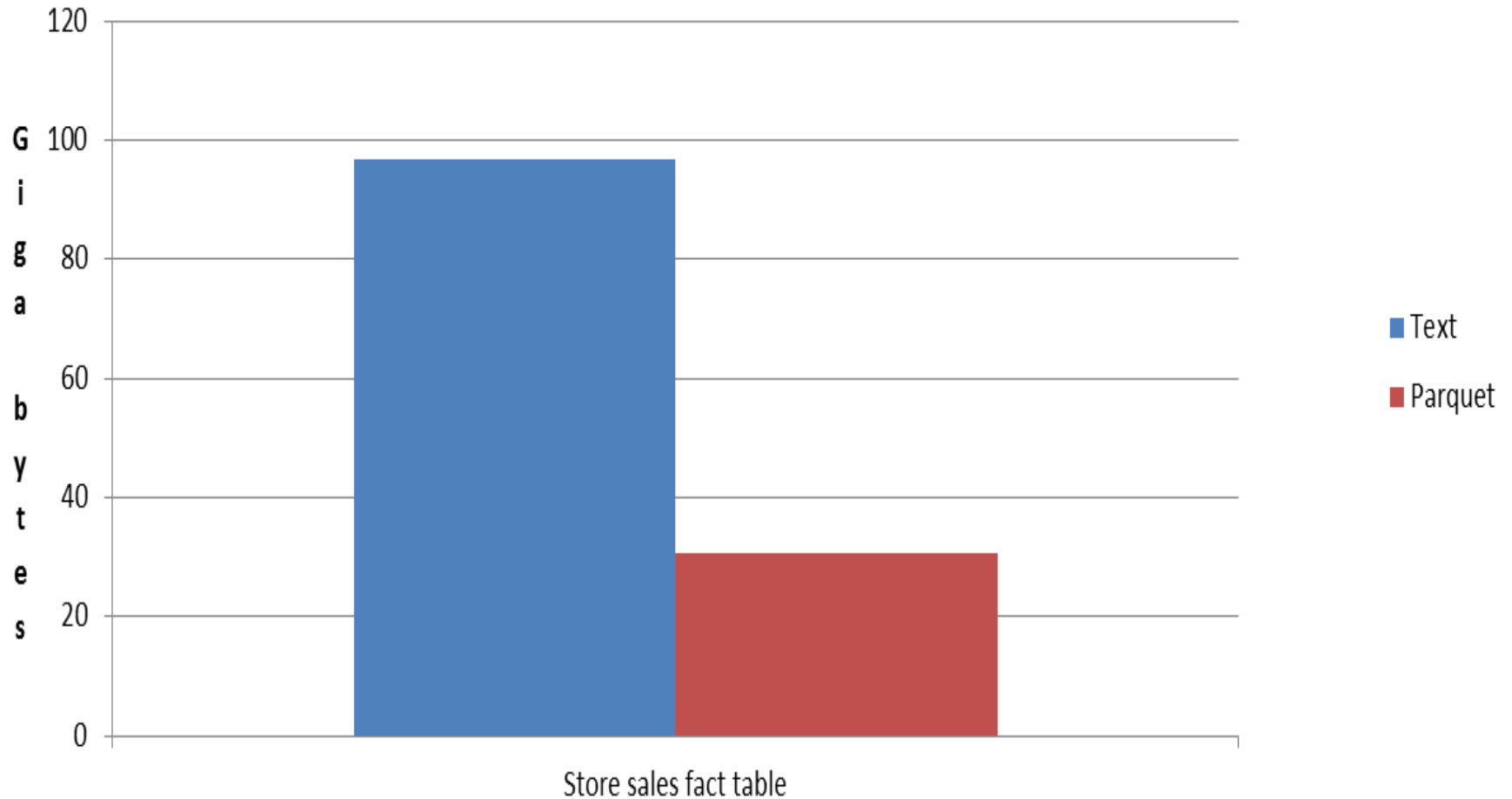
# Parquet design goals

- Interoperability
- Space efficiency
- Query efficiency



# Size comparison

TPCDS 100Gb scale factor



# Data Warehouse

Subject Area	Sales
Fact table size	100G
Sales transactions in Fact table	650 Million
Number of dimensions	9
Number of customers	12 Million

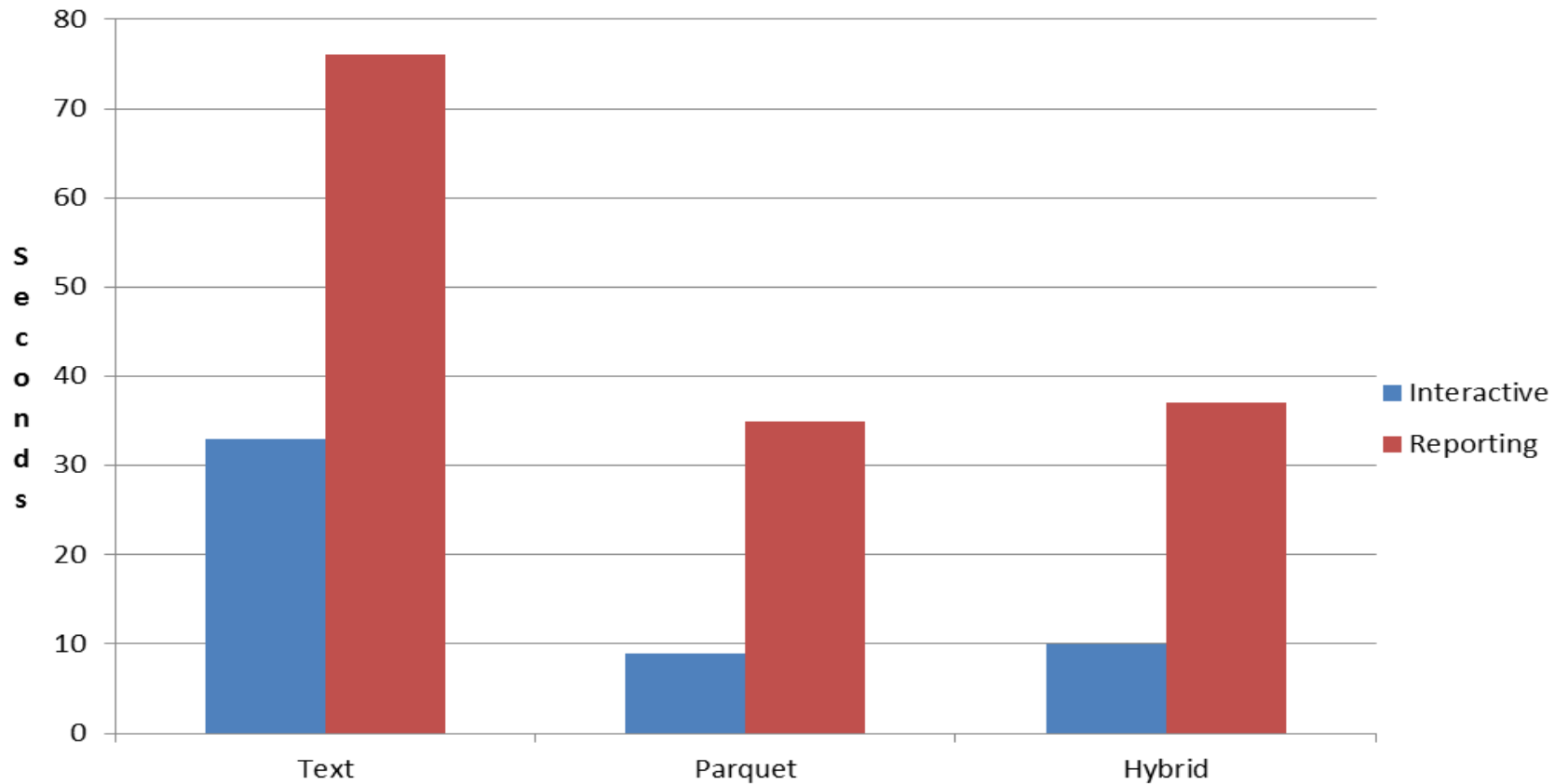
# Tools

- Cloudera's TPC-DS Toolkit
  - <https://github.com/cloudera/impala-tpcds-kit>



# Impala Query Performance

**4 machine  
32 cores  
240Gb Ram**




# Challenges

- Integrating Hive, Hbase with Impala
  - Documentation
- Conversion from text to parquet was very slow

## Work in Progress

- Tuning Impala, Hbase for performance

# About Me



**Data Architect Passionate about Engineering  
Data Warehouses and Data Lakes that drive  
Business Intelligence and Analytics**

**Extensive Experience in Capital Markets**

**Passionate about Nutrition, Yoga, Travel,  
Nature and Biking**