Augmented OLAP for Big Data

Li Yang
Kyligence CTO
Agenda

• About Kyligence
• Pains in Big Data Analysis
• Kyligence’s solution: Augmented OLAP
  • Video Demo
  • Benchmark
• Use Cases
Kyligence = Kylin + Intelligence

- Extreme multi-dimensional OLAP engine for big data
- Rank 1 from googling “big data OLAP”
- Rank 1 from googling “hadoop OLAP”
- 1000+ adoptions world wide
- Founded in 2016 by the team who created Apache Kylin
- CRN Top 10 Big Data Startups 2018
- Leading VCs: Redpoint Ventures, Cisco, CBC Capital and Shunwei Capital, Eight Roads Ventures (Fidelity International Arm), Coatue
Trusted by Fortune 500

Telecom
- China Telecom, #33 of Fortune 500
- China Mobile, #47 of Fortune 500
- China Unicom

Finance
- CPCIC, #252 of Fortune 500
- UnionPay
- China Merchants Bank
- Hua Tai Securities
- LU.com

Retail & Manufacture
- L'Oréal Paris, #392 of Fortune 500
- Huawei, #83 of Fortune 500
- Lenovo, #226 of Fortune 500
Global Partners

- Microsoft Global Gold Partner
- Amazon Web Service Technology Partner
- Tableau Technology Partner
- Cloudera Silver Partner
- MapR Converge Partner
- Hortonworks Community Partner
- Huawei Solution Partner
Agenda

• About Kyligence
• Pains in Big Data Analysis
• Kyligence’s solution: Augmented OLAP
• Use Cases
Why my report is so slow?

Fast and Changing Analysis Demand

VS

Slow and Heavy Big Data Operations
The Typical “Throw in some People” Approach

Business Analysis → Data Modeling → Lake → Warehouse → Mart → Reporting

High Cost: 🕒 $$$
**Pains in the “Throw in some People” Approach**

- **Time-to-value Pain**
  Weeks of waiting breaks the “online” promise.

- **Collaboration Pain**
  Hard to reuse asset across teams. Each team fights their own path.

- **Resource Pain**
  Hard to scale. Where to find so many skilled big data engineers?
Agenda

• About Kyligence
• Pains in Big Data Analysis
• Kyligence’s solution: Augmented OLAP
• Use Cases
Throw in some Intelligence!

Let a system replace the people.

- Transparent SQL Acceleration
- On-demand Data Preparation
- Interactive Query Performance
- High Concurrency
- Centralized Semantic Layer

Faster time to market. Stay “online”.
A Learning OLAP System

Business User → BI Tool

VS

Business User → Analyst → Data Engineer → Database → Graphs

Kyligence Inc. 2019, Confidential.
A Learning OLAP System

Business User → BI Tool

Patten Detection → Auto Modeling

Raw Data → Prepared Data

Augmented OLAP Engine

Data Preparing
Demo Setup

Analyze 1 billion rows of sales records (TPC-H)
Slow First Exploration
How to improve the first slow exploration?

What if the analyst operates differently the second time?

More comprehensive performance benchmark?
TPC-H Benchmark

- Examine large volumes of data
- High complexity queries
- Answers critical business questions
- 22 decision making queries

E.g. The Shipping Priority Query retrieves the shipping priority and potential revenue of the orders having the largest revenue among those that had not been shipped as of a given date. Top 10 orders are listed in decreasing order of revenue.
To see the trend as data grows

- 3 datasets
- Scale Factor = 20, 35, 50
Hardware Configurations

Same 4 physical nodes
- Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz * 2
- Totally 86 vCores, 188 GB mem

Same Spark configuration for both KE 4 Beta and SparkSQL 2.4
- spark.driver.memory=16g
- spark.executor.memory=8g
- spark.yarn.executor.memoryOverhead=2g
- spark.yarn.am.memory=1024m
- spark.executor.cores=5
- spark.executor.instances=17
Query Response Time | KE 4 Beta vs. SparkSQL 2.4

For each dataset:
- Run each query 3 times
- Record the average time
- No warm up

Lower is better.
Total Response Time | KE 4 Beta vs. SparkSQL 2.4

Total response time is the sum of 22 queries’ response time.

Compare over the size of datasets and feel the trend.

Scale out for the future.
Avg. Acceleration Rate | KE 4 Beta vs. SparkSQL 2.4

Acceleration Rate
\[\text{Acceleration Rate} = \frac{\text{SparkSQL time}}{\text{KE time}}\]

Take average of the 22 and compare over size of datasets.
Agenda

• About Kyligence
• Pains in Big Data Analysis
• Kyligence’s solution: Augmented OLAP
• Use Cases
A Successful Transition: China UnionPay

Augmented OLAP

Hive  Impala  Spark SQL  Drill
MapReduce  Spark  

Hive  Impala  Spark SQL  Drill
MapReduce  Spark  

Kyligence
Use Case: China UnionPay

Self-Service Big Data Warehouse
PB level (300B records)
big data warehouse of both self-service aggregation query and raw data query by business analysts

Merchant or Card Multi-dimensional Analytics
Support analysis on high granularity dimensions such as Merchant (10M cardinality) and Card (10B cardinality)

Efficient IT Operation
Significantly increase IT operation efficiency as 1 Kyligence cube replacing 800 Cognos cubes with unified data access management

More scalable Architecture
Kyligence scale-out architecture provide best flexibility for IT infrastructure when faced with increasing data and concurrent analysis demands
Use Case: China UnionPay

Card Transaction Analysis Portfolio

Functional Scene
- Org. Daily Cube
- Merch. Daily Cube
- Channel Daily Cube
- Region Daily Cube

Time Scene
- Org. Monthly Cube
- Merch. Monthly Cube
- Channel Monthly Cube
- Region Monthly Cube

Geo Scene
- Shanghai Merchants
- Zhejiang Merchants
- Anhui Merchants
- Guangdong Merchants

One Card Tx Model
- Dimensions: 167
- Measures: 20

800+ Cognos Cube, 1000+ ETL jobs
Thank You!

Homepage:  http://kyligence.io
Twitter:  @kyligence
Booth:  #1327