Apache Kylin Use Cases in China and Japan

Luke Han | luke@kyligence.io
VP of Apache Kylin, ASF Member
Co-founder & CEO of Kyligence Inc.
Agenda

- About Apache Kylin
- Why Apache Kylin
- Use Cases
- Q & A
About Kyligence Inc.

- Formed by the team who created Apache Kylin
- Funding from Redpoint Ventures, CBC and Shunwei Capital
- Offering Data Warehouse products and services powered by Apache Kylin

#StrataData
About Apache Kylin

--Kylin /ˈkiːlɪn/ 麒麟
--n. (in Chinese art) a mythical animal of composite form

--Apache Kylin
Extreme OLAP Engine for Big Data

✓ Leading open source OLAP on Hadoop
✓ Fast growing open source community
✓ Adopted by 200+ global organizations
✓ First born in China Apache Top Level Project
✓ InfoWorld Bossie Award:
  Best Open Source Big Data Tool (2015)
  Best Open Source Big Data Tool (2016)

http://kylin.apache.org
Apache Kylin: the missing part of Big Data

- High Performance
- High Concurrency
- ANSI SQL
- Native on Hadoop
- Cloud Ready
Global Users

Yahoo! Japan

200+ use cases in production
**OLAP Cube: Space for Time**

- Base vs. aggregate cells; ancestor vs. descendant cells; parent vs. child cells

1. (9/15, milk, Urbana, Dairy_land) - <time, item, location, supplier>
2. (9/15, milk, Urbana, *) - <time, item, location>
3. (*, milk, Urbana, *) - <item, location>
4. (*, milk, Chicago, *) - <item, location>
5. (*, milk, *, *) - <item>

**OLAP Cube**

- Cuboid = one combination of dimensions
- Cube = all combination of dimensions (all cuboids)
The Magic: Pre-Calculation

```
select returned, status, sum(quantity), sum(price)
from lineitem inner join orders on l_orderkey = o_orderkey
where shipdate <= '1998-09-16'
group by returned, status
order by returned, status;
```
Apache Kylin Architecture

1. Fetch source from Hive/Kafka
2. Build Cube with MR/Spark
3. Store Cube in HBase
4. Serve with ANSI SQL
Why Apache Kylin? Speed!

- Kylin vs Hive: O(1) vs O(n)
- Kylin’s query latency is better than Hive and others
- Lower is better

**Kylin vs Hive**

- SF: Scale Factor, SF 10 = 60 millions rows

**Kylin vs A popular SQL on Hadoop**

- Kylin is aim to speed up SQL analytics for OLAP, not aim to replace SQL on Hadoop

#StrataData
Why Apache Kylin? Stable!

Response Time

Data Size

Other Engines

O(N)

Apache Kylin

O(1)
Why Apache Kylin? Scale!

- Run Star-Schema Benchmark at 10, 20 and 40 million-row scales
- Kylin offers second latency when scale to billion even trillion data level

SF: Scale Factor, SF 10 = 60 millions rows
Apache Kylin Use Cases
Use Case – Analytics on Trillion Data

**TOUTIAO**: Top 1 news feed app in China

*Apache Kylin is the best OLAP on Hadoop solution, we can't do analytics on trillion level data without Apache Kylin*

--- Chaozhong Yang, Bytedance (Taotiao.com)

- Top 1 News Feed App
- 600+M Users, 8000+ DAU
- 70+ minutes /user/day

**Pain Points**
- Huge Volume Data, 100+B rows per day
- Second query latency by business requirement

**Solution**
- Apache Kylin as OLAP platform with 60+ HBase region servers

**Scenarios**
1. Video impression data, 4+ TB Cube (expansion rate < 3%)  
   *(Raw records: 3,000+B, 100+TB)*
2. APP performance monitoring, **1000+ Analysts**
3. News Log Insight, **100+ Cubes** for different products

**Benefits**
1. Super fast query speed: 90%+ queries in ms level  
   *(10,000+ times faster than hive)*
2. Saving Cluster Computing Resources: **One Build for all queries**
3. Rich API and drives, managing 100+Cube automatically
Use Case – OLAP on Hadoop

**Meituan**: Top 1 O2O company in China

### Challenge
- Slow performance with previous MySQL option
- Heavy development efforts with Hive solution
- Huge resources for Hive job
- Analysts can’t access directly for data on Hadoop

### Solution
- Apache Kylin as core OLAP on Hadoop solution
- SQL interface for internal users
- Active participate in open source Kylin community

**214 Cubes**

**500K queries/day**

**285.3+B Data Size**

**59TB Cube Storage**

**TP50 87ms**

**TP99 1266ms**

Supporting all critical business lines including E-Takeaways, Hotel, Movie…
Use Case – Migrating DW/BI to Big Data

CPIC: China Pacific Insurance (Group) Co., Ltd, Top3 insurance company in China

Challenge
Legacy DW/BI system is facing challenge for today’s need:
• IBM Cognos can’t fulfill requirement with huge volume data
• Billions data volume requires second query latency
• Hard for Analysts to learn Spark/MapReduce

Solution
• Built a big data analytics platform with Enterprise Kylin on top of Hadoop
• Using IBM Cognos as presentation layer only
• Analyst build data model via Enterprise KAP without IT

Benefits
The new platform enables analysts to:
• Continue DW/BI experts daily work with new tech stack without learning programing or other “high-tech”
• Supports national wide analytics
• Unleashed Big Data platform’s power with Apache Kylin
Use Case – Mobile OS Usage Analytics

**OPPO**: Top 4 Global Smartphone Vendor

The most important value of Apache Kylin (KAP) is it aggressively reduced efforts and duration to build a Data Warehouse on Hadoop which speeded up, simplified data access for Analysts, increase efficiency is the key for productivity.

-- Louis Wong

-- Head of Big Data ,OPPO

Built a Mobile OS Usage Analytics platform with Apache Kylin to replace Hive/Impala reporting system which has lack interactive analytics capability for their analysts group.

**With Apache Kylin (now KAP)**

- 100+B new row/day
- 90%+ query latency < 2 second
- Data preparation < 15 minutes
Use Case – Unified KPI Platform

**Home Link (Lianjia):** Top 1 O2O real estate agency service provider in China

**Challenge**
- Rapid business growth (5k stores in 2015, 8K stores in 2016) requires ultra-fast reporting system which could serve huge volume data
- Business require unified KPI definition and system

**Benefits**
- Unified KPI Platform for KPIs
- Managing critical KPIs in one place
- KPI number: 170+
- Cube Storage: 6+ TB
- Biggest Cube: 2+B rows
- 80% queries less than 1 second
Use Case – Shopping Reporting System

Yahoo! Japan: the most visited website in Japan

- We provide a reporting system that show statistics for store owners.
  - e.g. impressions, clicks and sales.
- Our reporting system used Impala as a backend database previously.
  - It took a long time (about 60 sec) to show Web UI.
- In order to lower the latency, we moved to Apache Kylin.
  - Average latency < 1sec for most cases

- Thanks to low latency with Kylin, we become possible to focus on adding functions for users.
Use Case – Kylin on Cloud

**Strikingly:** first Chinese company to graduate from the Y-Combinator seed accelerator.

- Built a Cloud Analytics Platform with AWS EMR & Apache Kylin to replace existing technology provided by cloud vendor.
  - Reduced query latency from 5~10 to less than 1 second
  - Bring high concurrency queries with stable performance (latency be guaranteed)
  - Elastic resizing computing cluster to save cost (only keep query cluster serving daily reporting needs)
More...

- **eBay**
  - Apache Kylin origins from eBay, various use cases from behavior, streaming, reporting and so on

- **JD.com Cloud**
  - Cloud deployment serves cloud usage and dashboard for merchant

- **Glispa**
  - using Apache Kylin as an OLAP component within its data management platform (DMP)

- **Exponential**
  - Using Kylin as a component in its data platform LAMBDA to power inventory, campaign, behavior and demand analysis for advertising

- **Powered By:**
  - [http://kylin.apache.org/community/poweredby.html](http://kylin.apache.org/community/poweredby.html)
Thanks

- For Apache Kylin
  - Please visit: http://kylin.apache.org

- For Kyligence Inc:
  - Please visit: http://kyligence.io

- Contact me:
  - luke@kyligence.io