# KylinProposal

## Apache Kylin

### Abstract

Kylin is a distributed and scalable OLAP engine built on Hadoop to support extremely large datasets.

### Proposal

Kylin is an open source Distributed Analytics Engine that provides multi-dimensional analysis (MOLAP) on Hadoop. Kylin is designed to accelerate analytics on Hadoop by allowing the use of SQL-compatible tools. Kylin provides a SQL interface and multi-dimensional analysis (MOLAP) on Hadoop to support extremely large datasets and tightly integrate with Hadoop ecosystem.

#### Overview of Kylin

Kylin platform has two parts of data processing and interactive: First, Kylin will read data from source, Hive, and run a set of tasks including Map Reduce job, shell script to pre-calcuate results for a specified data model, then save the resulting OLAP cube into storage such as HBase. Once these OLAP cubes are ready, a user can submit a request from any SQL-based tool or third party applications to Kylin's REST server. The Server calls the Query Engine to determine if the target dataset already exists. If so, the engine directly accesses the target data in the form of a predefined cube, and returns the result with sub-second latency. Otherwise, the engine is designed to route non-matching queries to whichever SQL on Hadoop tool is already available on a Hadoop cluster, such as Hive.

Kylin platform includes:

**Metadata Manager:** Kylin is a metadata-driven application. The Kylin Metadata Manager is the key component that manages all metadata stored in Kylin including all cube metadata. All other components rely on the Metadata Manager.

**Job Engine:** This engine is designed to handle all of the offline jobs including shell script, Java API, and Map Reduce jobs. The Job Engine manages and coordinates all of the jobs in Kylin to make sure each job executes and handles failures.

**Storage Engine:** This engine manages the underlying storage – specifically, the cuboids, which are stored as key-value pairs. The Storage Engine uses HBase – the best solution from the Hadoop ecosystem for leveraging an existing K-V system. Kylin can also be extended to support other K-V systems, such as Redis.

**Query Engine:** Once the cube is ready, the Query Engine can receive and parse user queries. It then interacts with other components to return the results to the user.

**REST Server:** The REST Server is an entry point for applications to develop against Kylin. Applications can submit queries, get results, trigger cube build jobs, get metadata, get user privileges, and so on.

**ODBC Driver:** To support third-party tools and applications – such as Tableau – we have built and open-sourced an ODBC Driver. The goal is to make it easy for users to onboard.

### Background

The challenge we face at eBay is that our data volume is becoming bigger and bigger while our user base is becoming more diverse. For e.g. our business users and analysts consistently ask for minimal latency when visualizing data on Tableau and Excel. So, we worked closely with our internal analyst community and outlined the product requirements for Kylin:

1. Sub-second query latency on billions of rows 2. ANSI SQL availability for those using SQL-compatible tools 3. Full OLAP capability to offer advanced functionality 4. Support for high cardinality and very large dimensions 5. High concurrency for thousands of users 6. Distributed and scale-out architecture for analysis in the TB to PB size range

Existing SQL-on-Hadoop solutions commonly need to perform partial or full table or file scans to compute the results of queries. The cost of these large data scans can make many queries very slow (more than a minute). The core idea of MOLAP (multi-dimensional OLAP) is to pre-compute data along dimensions of interest and store resulting aggregates as a "cube". MOLAP is much faster but is inflexible.
We realized that no existing product met our exact requirements externally – especially in the open source Hadoop community. To meet our emerging business needs, we built a platform from scratch to support MOLAP for these business requirements and then to support more others include ROLAP. With an excellent development team and several pilot customers, we have been able to bring the Kylin platform into production as well as open source it.

### Rationale

When data grows to petabyte scale, the process of pre-calculation of a query takes a long time and costly and powerful hardware. However, with the benefit of Hadoop's distributed computing architecture, jobs can leverage hundreds or thousands of Hadoop data nodes. There still exists a big gap between the growing volume of data and interactive analytics:

1. Existing Business Intelligence (OLAP) platforms cannot scale out to support fast growing data. 2. Existing SQL on Hadoop projects are not designed for OLAP use cases, huge tables joins will always take long time to scan and calculate. 3. No mature OLAP solution exists on Hadoop

As mentioned in the background, the business requirements triggered by increase in data volume drove eBay to invest in building a solution from scratch to offer Analytics capability on Hadoop cluster. With Hadoop's power of distributed computing Kylin can perform pre-calculations in parallel and merge the final results, thereby significantly reducing the processing time.

To serve queries by the analyst community, Kylin generates cuboids with all possible combinations of dimensions, and calculate all metrics at different levels. The cuboids are then integrated to form a pre-calculated OLAP cube. All cuboids are key-value structured: keys are composites formed from combinations of multiple dimensions and values are aggregations results for that particular combination of dimensions. Kylin uses HBase to store cubes. HBase is useful because it supports efficient searches across ranges of data.

# Current Status

## Meritocracy

Kylin has been deployed in production at eBay and is processing extremely large datasets. The platform has demonstrated great performance benefits and has proved to be a better way for analysts to leverage data on Hadoop with a more convenient approach using their favorite tool.

## Community

Kylin seeks to develop developer and user communities during incubation.

## Core Developers

Kylin is currently being designed and developed by six engineers from eBay Inc. – Jiang Xu, Luke Han, Yang Li, George Song, Hongbin Ma and Xiaodong Duo. In addition, some outside contributors are actively contributing in design and development. Among them, Julian Hyde from Hortonworks is a very important contributor. All of these core developers have deep expertise in Hadoop and the Hadoop Ecosystem in general.

## Alignment

The ASF is a natural host for Kylin given that it is already the home of Hadoop, Pig, Hive, and other emerging cloud software projects. Kylin was designed to offer OLAP capability on Hadoop from the beginning in order to solve data access and analysis challenges in Hadoop clusters. Kylin complements the existing Hadoop analytics area by providing a comprehensive solution based on pre-computed views.

In Kylin, we are leveraging an open-source dynamic data management framework called Apache Calcite to parse SQL and plug in our code. Apache Calcite was previously called Optiq, was originally authored by Julian Hyde and is now an Apache Incubator project.

# Known Risks

## Orphaned Products

The core developers of Kylin team plan to work full time on this project. There is very little risk of Kylin getting orphaned since at least one large company (eBay) is extensively using it in their production Hadoop clusters. For example, currently there are 3 use cases with more that 12+Billion rows and 1000 activity requests per day using Kylin in production. Furthermore, since Kylin was open sourced at the beginning of October 2014, it has received more than 280 stars and been forked nearly 100 times. Kylin has one major release so far and and received 5 pull requests from contributors in the first month pull requests from external sources in the last month, which further demonstrates Kylin as a very active project. We plan to extend and diversify this community further through Apache.

## Inexperience with Open Source

The core developers are all active users and followers of open source. They are already committers and contributors to the Kylin Github project. All have been involved with the source code that has been released under an open source license, and several of them also have experience developing code in an open source environment. Though the core set of Developers do not have Apache Open Source experience, there are plans to onboard individuals with Apache open source experience on to the project.

## Homogenous Developers

The core developers include developers from eBay, Ctrip and Hortonworks. Apache Incubation process encourages an open and diverse meritocratic community. Apache Kylin has the required amount of diversity with committers from three different organizations, but is also aware that bulk of the commits come from a single entity. Kylin intends to make every possible effort to build a diverse, vibrant and involved community and has already received substantial interest from various organizations

## Reliance on Salaried Developers

eBay invested in Kylin as the OLAP solution on top of Hadoop clusters and some of its key engineers are working full time on the project. In addition, since there is a growing Big Data need for scalable OLAP solutions on Hadoop, we look forward to other Apache developers and researchers to contribute to the project. Additional contributors, including Apache committers have plans to join this effort shortly. Also key to addressing the risk associated with relying on Salaried developers from a single entity is to increase the diversity of the contributors and actively lobby for Domain experts in the BI space to contribute. Apache Kylin intends to do this. One approach already taken is to approach the Apache Drill project to explore possible cooperation.

## Relationships with Other Apache Products

Kylin has a strong relationship and dependency with Apache Hadoop HBase, Hive and Calcite. Being part of Apache's Incubation community, could help with a closer collaboration among these four projects and as well as others.

Kylin is likely to have substantial value to Apache Drill due to the common use of Calcite as a query optimization engine and similar approaches between Kylin's approach to cubing and Drill's approach to input sources.

### An Excessive Fascination with the Apache Brand

Kylin is proposing to enter incubation at Apache in order to help efforts to diversify the committer-base, not so much to capitalize on the Apache brand. The Kylin project is in production use already inside EBay, but is not expected to be an EBay product for external customers. As such, the Kylin project is not seeking to use the Apache brand as a marketing tool.

## Documentation

Information about Kylin can be found at https://github.com/KylinOLAP/Kylin. The following links provide more information about Kylin in open source:

- Kylin web site: http://kylin.io
- Codebase at Github: https://github.com/KylinOLAP/Kylin
- Issue Tracking: https://github.com/KylinOLAP/Kylin/issues
- User community: https://groups.google.com/forum/#!forum/kylin-olap

## Initial Source

Kylin has been under development since 2013 by a team of engineers at eBay Inc. It is currently hosted on Github.com under an Apache license at https://github.com/KylinOLAP/Kylin

## External Dependencies

Kylin has the following external dependencies.

#### Basic

- JDK 1.6+
- Apache Maven
- JUnit
- DBUnit
- Log4j
- Slf4j
- Apache Commons
- Google Guava
- Jackson

#### Hadoop

- Apache Hadoop
- Apache HBase
- Apache Hive
- Apache Zookeeper
- Apache Curator

#### Utility

- H2
- JSCH

#### REST Service

- Spring

#### Query

- Antlr
- Apache Calcite (formerly Optiq)
- Linq4j

#### Job

- Quartz

#### Web build tool

- NPM
- Grunt
- bower

### Web

- Angular JS
- jQuery
- Bootstrap
- D3 JS
- ACE

## Cryptography

Kylin will eventually support encryption on the wire. This is not one of the initial goals, and we do not expect Kylin to be a controlled export item due to the use of encryption. Kylin supports but does not require the Kerberos authentication mechanism to access secured Hadoop services.

## Required Resources

### Mailing List

- kylin-private for private PMC discussions (with moderated subscriptions)
- kylin-dev
- kylin-commits

### Subversion Directory

Git is the preferred source control system: git://git.apache.org/Kylin

### Issue Tracking

JIRA Kylin (KYLIN)

### Other Resources

The existing code already has unit tests so we will make use of existing Apache continuous testing infrastructure. The resulting load should not be very large.

## Initial Committers

- Jiang Xu < jiangxu.china at gmail dot com>
- Luke Han <lukhan at ebay dot com>
- Yang Li <yangli9 at ebay dot com>
- George Song <ysong1 at ebay dot com>
- Hongbin Ma <honma at ebay dot com>
- Xiaodong Duo < oranjedog at gmail dot com>
- Julian Hyde < jhyde at apache dot org >
- Ankur Bansal < abansal at ebay dot com>

## Affiliations

The initial committers are employees of eBay Inc., Ctrip and Hortonworks. The nominated mentors are employees of Hortonworks, MapR Technologies and Pivotal.

## Sponsors

### Champion

- Owen O'Malley < omalley at apache dot org >
- Ted Dunning <tdunning at apache dot org>

### Nominated Mentors

- Owen O'Malley < omalley at apache dot org > - Apache IPMC member, Co-founder and Senior Architect, Hortonworks
- Ted Dunning < tdunning at apache dot org> - Apache IPMC member, Chief Architect, MapR Technologies
- Henry Saputra <hsaputra at apache dot org> - Apache IPMC member, Pivotal
- Jacques Nadeau <jacques at apache dot org> (pending admission to IPMC) - Apache Drill PMC Chair, MapR Technologies

## Sponsoring Entity

We are requesting the Incubator to sponsor this project.