# LensProposal

## Lens

### Abstract

Lens is a platform that enables multi-dimensional queries in a unified way over datasets stored in multiple warehouses. Lens integrates Apache Hive with other data warehouses by tiering them together to form logical data cubes.

### Proposal

Lens provides a unified Cube abstraction for data stored in different stores. Lens tiers multiple data warehouses for unified representation and efficient access. It provides SQL-like Cube query language to query and describe data sets organized in data cubes. It enables users to run queries against Facts and Dimensions that can span multiple physical tables stored in different stores.

The primary use cases that Lens aims to solve:

- Facilitate analytical queries by providing the OLAP like Cube abstraction
- Data Discovery by providing single metadata layer for data stored in different stores
- Unified access to data by integrating Hive with other traditional data warehouses

### Background

Apache Hive is a data warehouse that facilitates querying and managing large datasets stored in distributed storage systems like HDFS. It provides SQL like language called HiveQL aka HQL. Apache Hive is a widely used platform in various organizations for doing adhoc analytical queries. In a typical Data warehouse scenario, the data is multi-dimensional and organized into Facts and Dimensions to form Data Cubes. Lens provides this logical layer to enable querying and manage data as Cubes. The Lens project is actively being developed at InMobi to provide the higher level of analytical abstraction to query data stored in different storages including Hive and beyond seamlessly.

### Rationale

The Lens project aims to ease the analytical querying capabilities and cut the data-silos by providing a single view of data across multiple data stores. Conceiving data as a cube with hierarchical dimensions leads to conceptually straightforward operations to facilitate analysis. Integrating Apache Hive with other traditional warehouses provides the opportunity to optimize on the query execution cost by tiering the data across multiple warehouses. Lens provides

- Access to data Cubes via Cube Query language similar to HiveQL.
- Driver based architecture to allow for plugging systems like Hive and other warehouses such as columnar data RDBMS.
- Cost based engine selection that provides optimal use of resources by selecting the best execution engine for a given query.

In a typical Data warehouse, data is organized in Cubes with multiple dimensions and measures. This facilitates the analysis by conceiving the data in terms of Facts and Dimensions instead of physical tables. Lens aims to provide this logical Cube abstraction on Data warehouses like Hive and other traditional warehouses.

### Initial Goals

- Donate the Lens source code and documentation to Apache Software Foundation
- Build a user and developer community
- Support Hive and other Columnar data warehouses
- Support full query life cycle management
- Add authentication for querying cubes
- Provide detailed query statistics

### Long Term Goals

Here are some longer-term capabilities that would be added to Lens

- Add authorization for managing and querying Cubes
- Provide REST and CLI for full Admin controls
- Capability to schedule queries
- Query caching
- Integrate with Apache Spark. Creating Spark RDD from Lens query
- Integrate with Apache Optiq

### Current Status

The project is actively developed at InMobi. The first version is deployed at InMobi 4 months back. This version allows querying dimension and fact data stored in Hive over CLI. The source code and documentation is hosted at GitHub.

### Meritocracy

We intend to build a diverse developer and user community for the project following the Apache meritocracy model. We want to encourage contributors from multiple organizations, provide plenty of support to new developers and welcome them to be committers.

### Community

Currently the project is being developed at InMobi. We hope to extend our contributor and user base significantly in the future and build a solid open source community around Lens. Core Developers Lens is currently being developed by Amareshwari Sriramadasu, Sharad Agarwal and Jaideep Dhok from InMobi, and Sreekanth Ramakrishnan who is currently employed by SoftwareAG. Raghavendra Singh from InMobi has built the QA automation for Lens.

### Alignment

The ASF is a natural home to Lens as it is for Apache Hadoop, Apache Hive, Apache Spark and other emerging projects in Big Data space. We believe in any enterprise, multiple data warehouses will co-exist, as not all workloads are cost effective to run on single one. Apache Hive is one of the crucial data warehouse along with upcoming projects like Apache Spark in Hadoop ecosystem. Lens will benefit in working in close proximity with these projects. The traditional Columnar data warehouses complement Apache Hive as certain workloads continue to be cost effective to run in traditional columnar data warehouses. Having multiple data warehouses leads to data silos that Lens aims to cut within the enterprise and provide a holistic unified access to data.

## Known Risks

### Orphaned products & Reliance on Salaried Developers

There is little risk of Lens getting orphaned, as Lens is key part of the Data Platform stack at InMobi. The core Lens developers plan to work on it full-time. We think Lens will bring value in the Big Data space and we plan to grow the community of users and contributors.

### Inexperience with Open Source

All the core developers have long and significant experience in Apache projects and Hadoop ecosystem. Amareshwari Sriramadasu has long standing contributions to Apache Hadoop MapReduce and Apache Hive, she being PMC member of Hadoop and a committer of Hive. Sharad Agarwal is a PMC member of Hadoop and contributed to Hadoop YARN and Hadoop MapReduce. Srikanth Sundarrajan is a PMC member of Apache Falcon. Sreekanth Ramakrishnan is committer of Apache Hadoop. Jaideep Dhok has contributed patches to Apache Hive. Gunther is a PMC member of Apache Hive. Vikram is a committer of Apache Hive.

### Homogeneous Developers

The initial developers are employed by Hortonworks, InMobi and SoftwareAG. We are committed to recruiting additional committers from other companies based on their contribution to the project.

### Reliance on Salaried Developers

The majority of initial committers are paid by their employee to contribute to the project and few are contributing in their spare time. Once the project has a community built, we are committed to recruit committers and developers from outside the current core developers.

### Relationships with Other Apache Products

Lens is deeply integrated with other Apache projects. Lens uses and extends Apache Hive HCatalog to store and manage the Data cubes. It uses HDFS and Hive session management libraries. Lens has the driver-based architecture that allows for adding multiple execution drivers. Apart from integrating Apache Hive, it can be integrated with Apache Spark over Spark SQL or Shark, Apache Drill, Apache Tajo and Apache Phoenix. In future we want to use Apache Optiq in Lens for query optimization and cost based driver selection.

### An Excessive Fascination with the Apache Brand

The project is conceived from beginning to be in line with the Apache philosophy. As the core developers have good experience with Apache, the source code organization, build, review and commit process are highly influenced by Apache. We believe that Apache will be a solid home for Lens to grow and build the open source community. We have also described the reasons in the Rationale and Alignment sections.

## Documentation

http://inmobi.github.io/grill/

## Initial Source

The source is currently in github repository at: https://github.com/inmobi/grill

## Source and Intellectual Property Submission Plan

The complete Lens code is already under Apache Software License 2.

## External Dependencies

The dependencies all have Apache compatible licenses. These include Apache 2.0, BSD, MIT, EPL and CDDL licensed dependencies.

## Cryptography

None

## Required Resources

### Mailing lists

- lens-dev AT incubator DOT apache DOT org
- lens-commits AT incubator DOT apache DOT org
- lens-private AT incubator DOT apache DOT org

### Subversion Directory

Git is the preferred source control system: git://git.apache.org/incubator-lens

### Issue Tracking

JIRA Lens (LENS)

## Initial Committers

- Amareshwari Sriramadasu (amareshwari AT apache DOT org)
- Gunther Hagleitner (gunther AT apache DOT org)
- Jaideep Dhok (jaideep.dhok AT Inmobi DOT com)
- Raghavendra Singh (raghavendra.singh AT Inmobi DOT com)
- Sharad Agarwal (sharad AT apache DOT org)
- Sreekanth Ramakrishnan (sreekanth AT apache DOT org)
- Srikanth Sundarrajan (sriksun AT apache DOT org)
- Suma Shivaprasad (suma.shivaprasad AT Inmobi DOT com)
- Vikram Dixit (vikram AT apache DOT org)

## Affiliations

- Amareshwari SR (InMobi)
- Gunther Hagleitner (Hortonworks)
- Jaideep Dhok (InMobi)
- Raghavendra Singh (InMobi)
- Sharad Agarwal (InMobi)
- Sreekanth Ramakrishnan (SoftwareAG)
- Srikanth Sundarrajan (InMobi)
- Suma Shivaprasad (InMobi)
- Vikram Dixit (Hortonworks)

## Sponsors

### Champion

Vinod K <vinodkv AT apache DOT org> (Apache Member)

### Nominated Mentors

- Chris Douglas (Microsoft)
- Jacob Homan (Microsoft)
- Jean Baptiste Onofre (Talend)

- Vinod K (Hortonworks)

## Sponsoring Entity

Incubator PMC