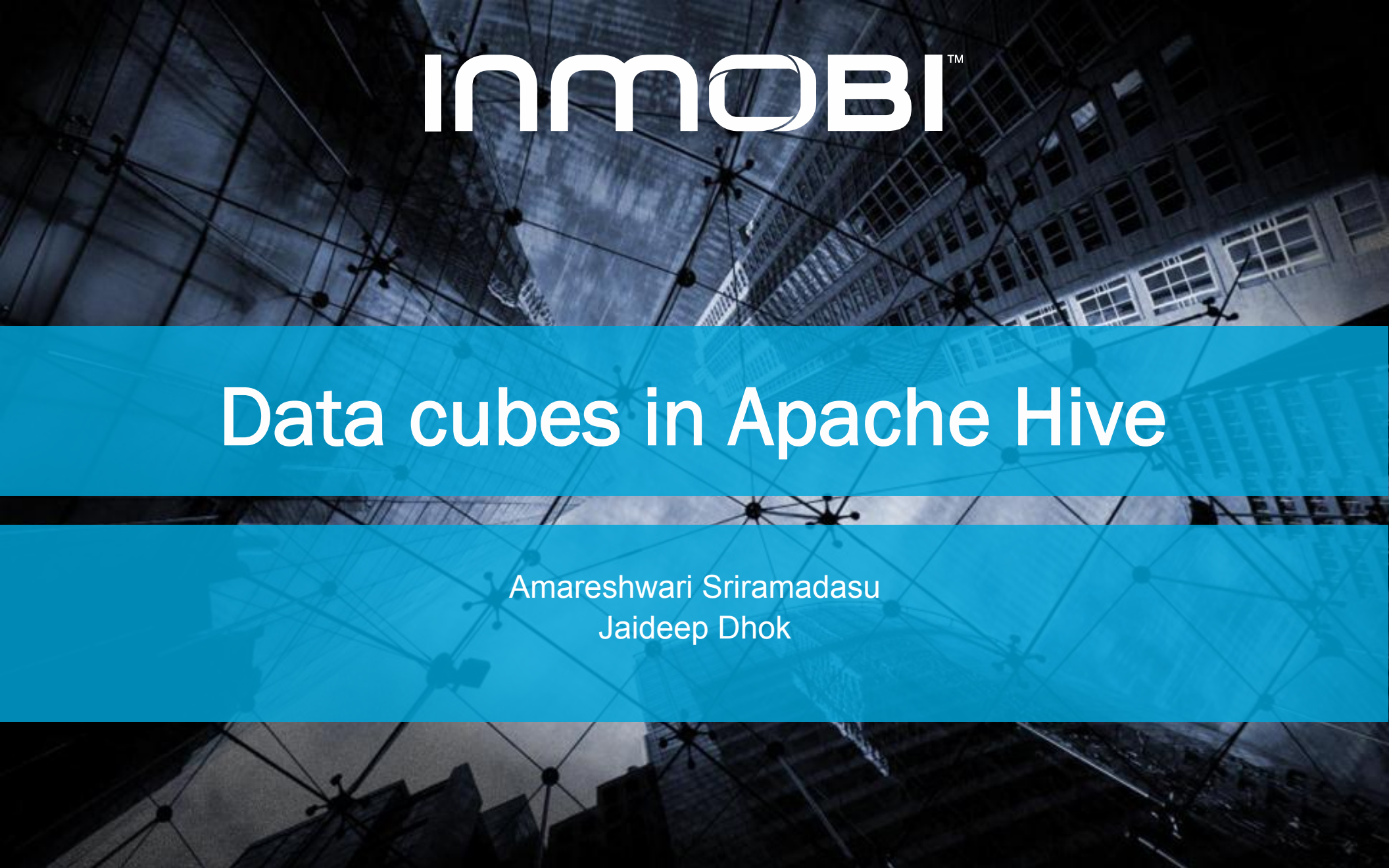


Data cubes in Apache Hive

Amareshwari Sriramadasu
Jaideep Dhok

Amareshwari Sriramadasu

- Engineer at Inmobi
- Apache Hive– Committer
- Apache Hadoop – PMC
- Working in Hadoop and eco systems since 2007

INMOBI™

Jaideep Dhok

- Engineer at Inmobi
- Working in distributed systems and Hadoop eco system since 2007
- Apache Hive and Hadoop Contributor

INMOBI™

Agenda

- Background
- Why Apache Hive
- Data cubes in Hive
- Queries on data cubes with examples
- Road map

Agenda

- Background
  - Why Apache Hive
  - Data cubes in Hive
  - Queries on data cubes with examples
- Road map

INMOBI™

About InMobi

Global Mobile technology company enabling

- Developers & Publishers to monetize
- Advertisers to engage and acquire users

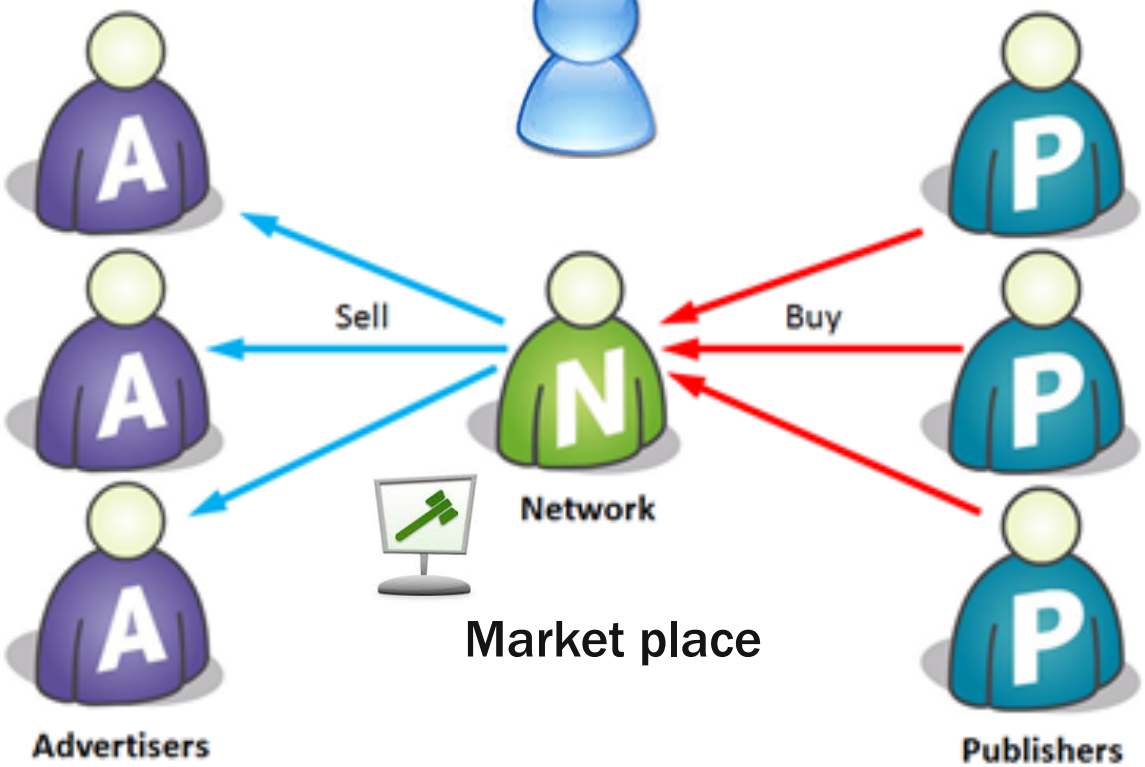
@ Scale

Digital advertising – Intro

Consumer

Wants to target Users

Brings money



Market place

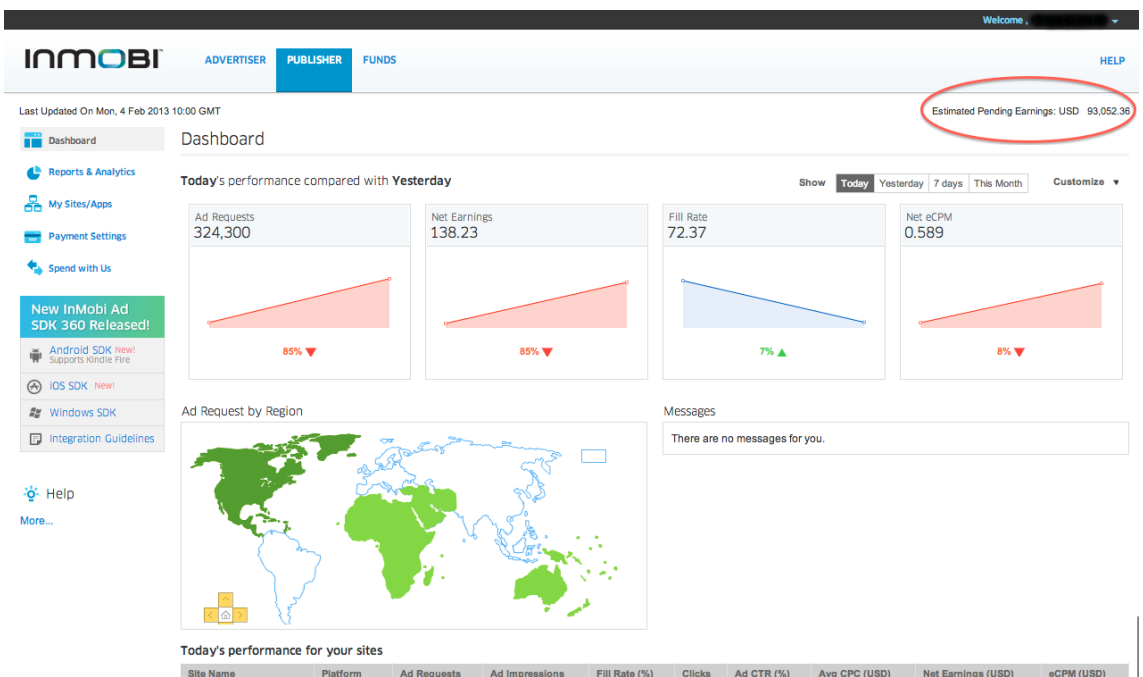
Owns & Sells Real estate on digital inventory

Has reach to users

Courtesy: http://www.liesdamnedlies.com/

INMOBI™

Hadoop @ InMobi – Factual Reporting & Analytics



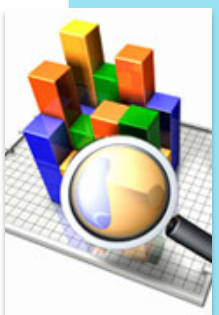
- 130 TB Hadoop warehouse
- 5 TB SQL warehouse
- Pipelines

INMOBI™

Use cases

Use cases

- Sharing reports/data with customer (account level)
- Understand trends through data & exploration (analysis)
- Debugging / Postmortem of issues (troubleshooting)
- Sizing & Estimation (Ex: inventory, reach)
- Summary of Product lines, Geographies, Network (Ex: Rev by Geo)
- Sales/Revenue Targets vs Actuals
- Tracking campaign performance
- Tracking any metrics on REAL- TIME basis



Categorize use cases

- Batch queries
- Adhoc queries
- Interactive queries
- Canned reports
- Scheduled reports
- Infer insights through ML algorithms

InMobi™

Current state of analytics - Reporting

Adhoc Querying system

- Adhoc and Batch queries
- Scheduled queries
- Based on Hadoop Mapreduce
- Provides UI and custom api
- Data is stored in HDFS

Dashboard system

- Canned reports
- Scheduled reports
- Interactive and adhoc queries
- Provides UI and Custom api
- Data is stored in columnar DWH, InfoBright

Customer facing system

- Face to the outside world (Advertisers and publishers)
- Interactive and adhoc queries
- Provides UI and custom api
- Data is stored in relational DB, Postgres

INMOBI™

Current State - Problems

- Disparate user experience
- Disparate data storage systems causing inability to scale
- Not leveraging community around

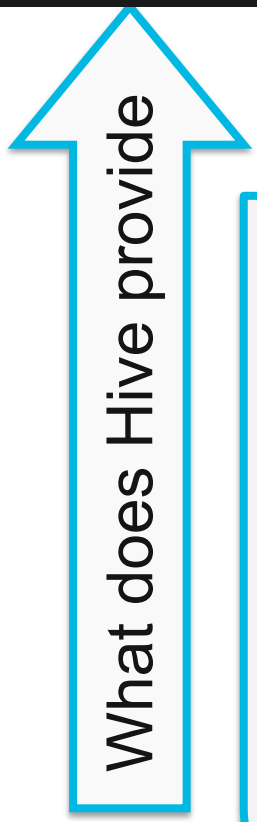
INMOBI™

Agenda

- Background
- Why Apache Hive
- Data cubes in Hive
- Queries on data cubes with examples
- Road map

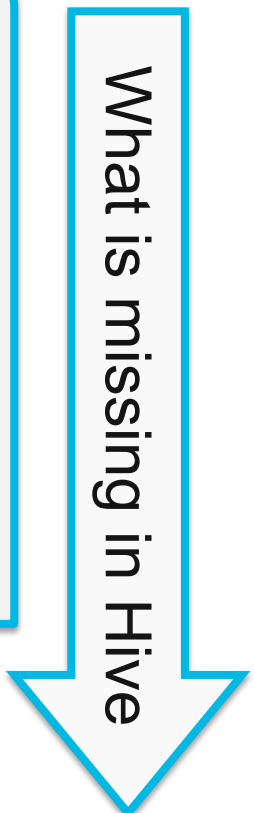
INMOBI™

Apache Hive to the rescue



- Associates structure to data
- Provides Metastore and catalog service – Hcatalog
- Provides pluggable storage
- Accepts SQL like queries
- HQL is widely adopted language by systems like Shark, Impala
- Provides pluggable interface for adding new storage
- Has strong apache community

- Data warehouse features like facts, dimensions
- Logical table associated with multiple physical storages
- Pluggable execution engine for HQL
- Query history, caching
- Scheduling queries



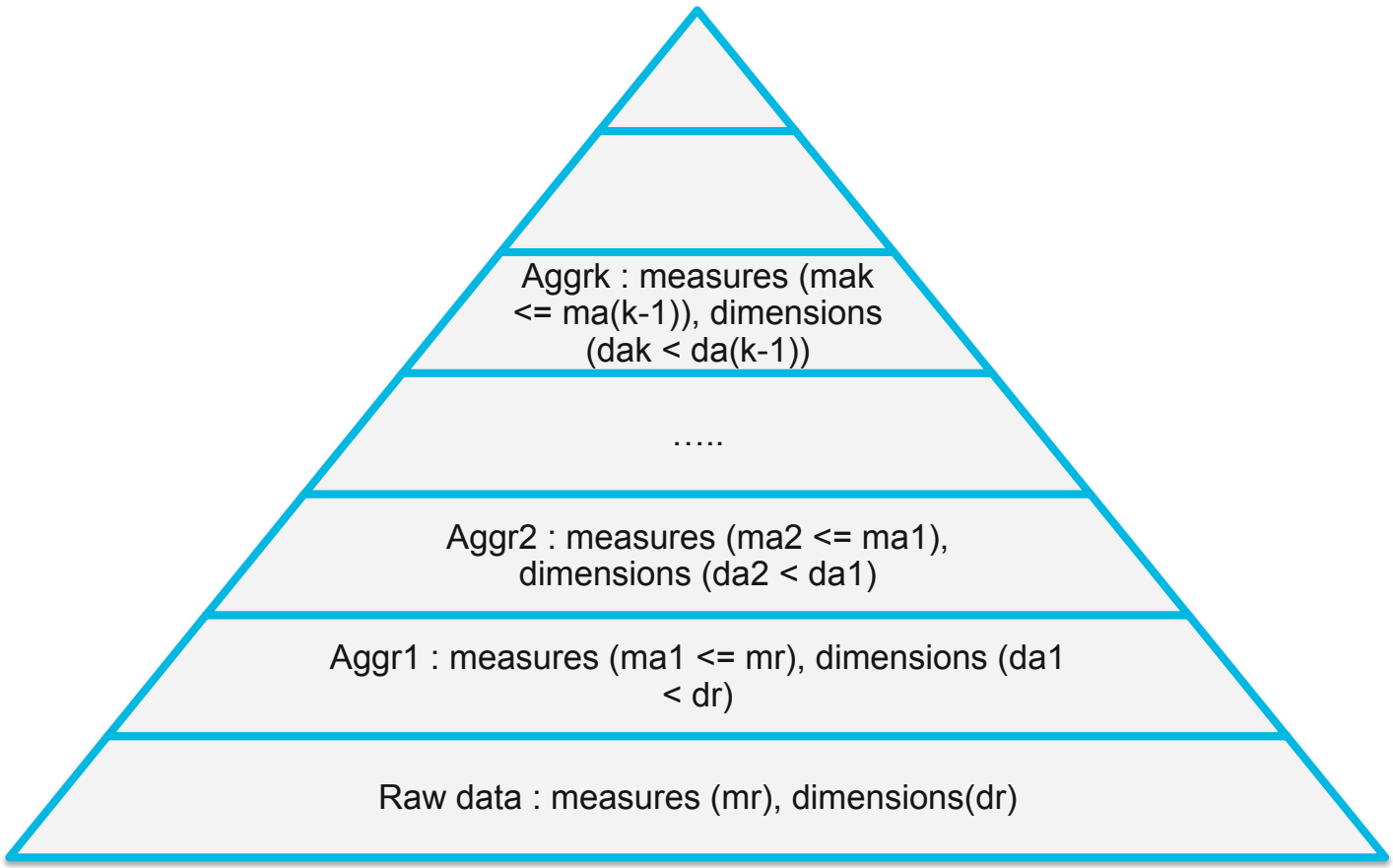
INMOBI™

Agenda

- Background
- Why Apache Hive
- Data cubes in Hive
- Queries on data cubes with examples
- Road map

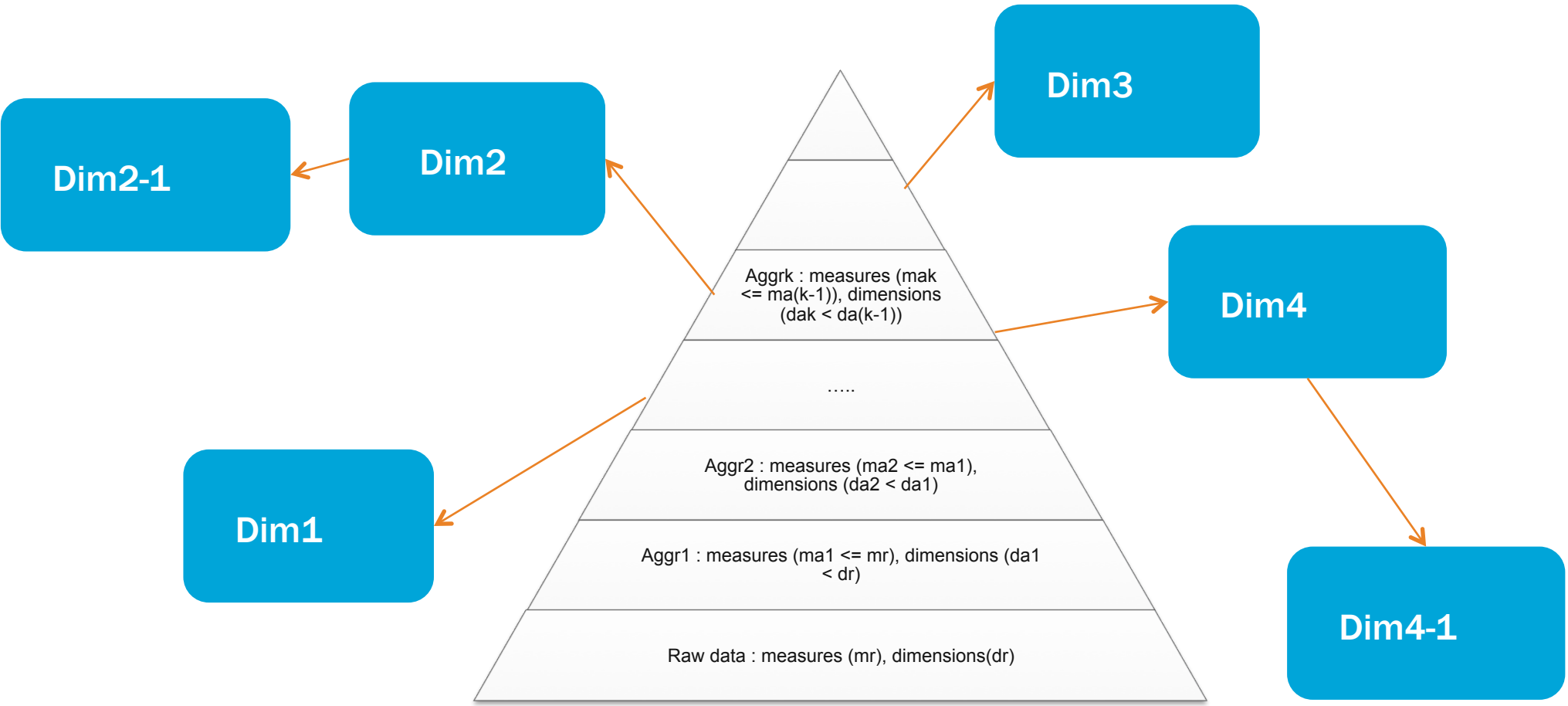
INMOBI™

Data Layout



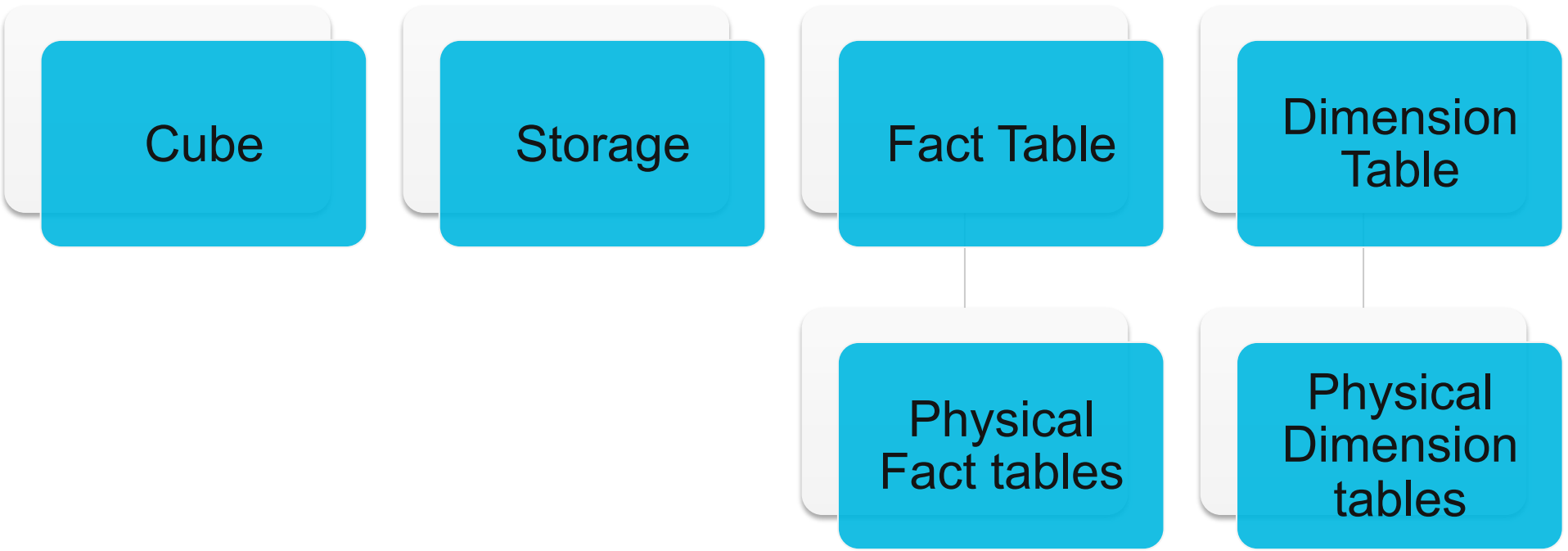
Other side of pyramid is aggregated at timed dimension

Data Layout



INMOBI™

Data Model



InMobi

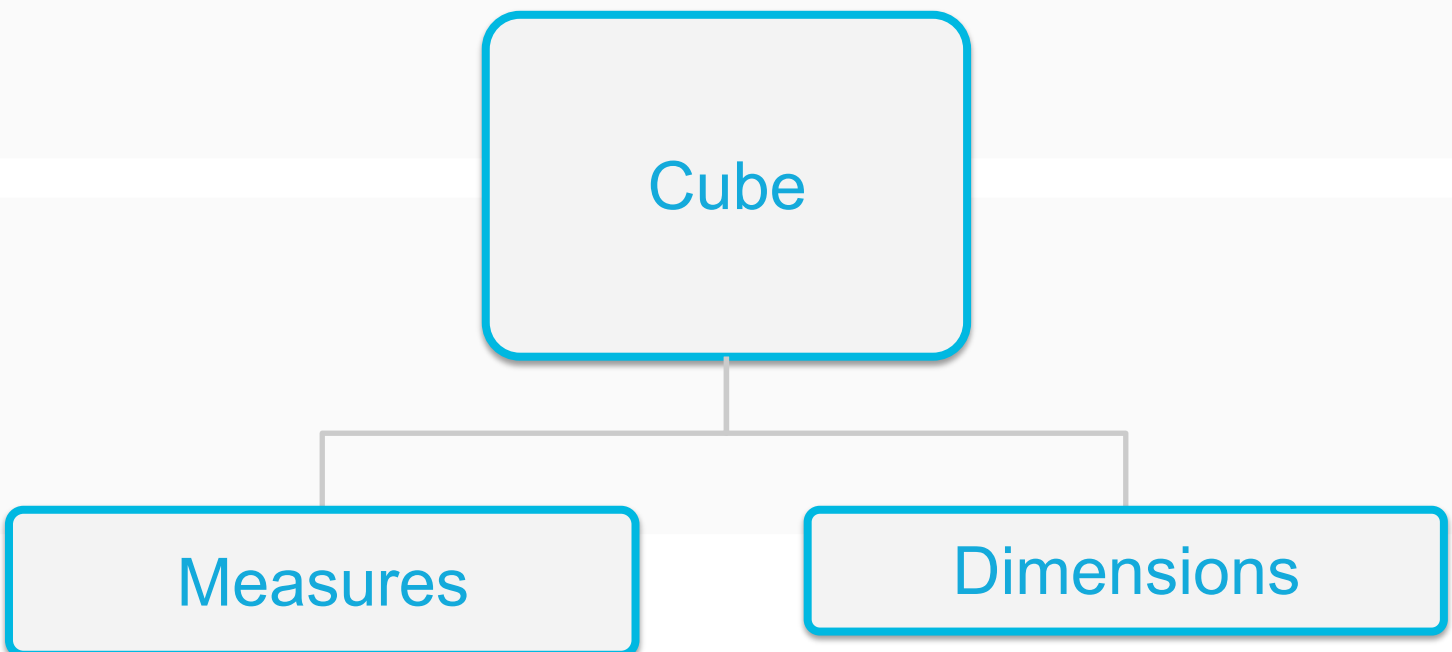
Data Model - Cube

Measure

- Column Measure : name, type, default aggregate, format string, start date, end date
- Expression Measure : Associated Expression

Dimension

- Simple Dimension: name, type, start date, end date
- Referenced Dimension : Referencing table and column
- Hierarchical Dimension :hierarchy
- Expression Dimension : Associated expression



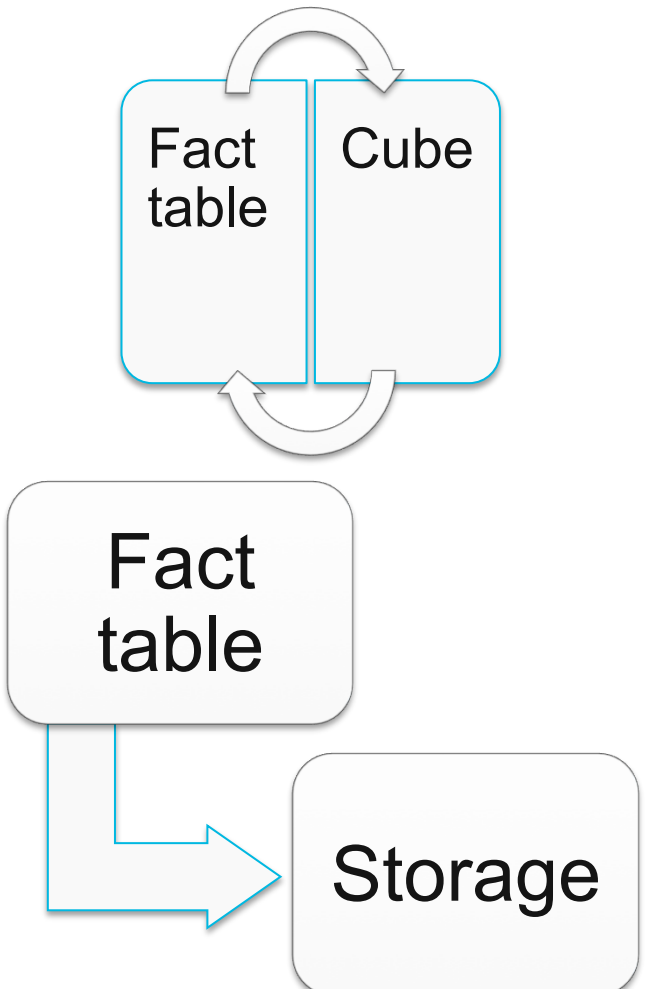
INMOBI™

Storage

- Name
- End point
- Properties
- Ex : UA2, UJ1, Mpower-IB

Fact Table

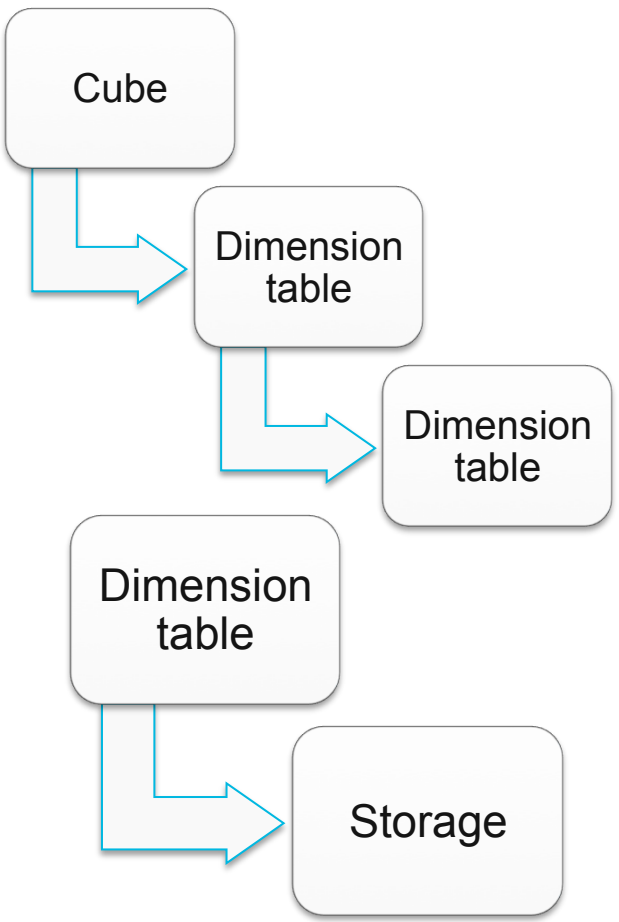
- Columns
- Cube that it belongs
- Storages on which it is present and the associated update periods



INMOBI™

Dimension Table

- Columns
- Dimension references
- Storages on which it is present and associated snapshot dump period, if any.



Data Model – Storage tables and partitions

Storage table

- Associated storage descriptor
- Partitioned by columns

Fact table

- Fact storage table

Dimension table

- Dimension storage table

INMOBI™

Agenda

- Background
- Why Apache Hive
- Data cubes in Hive
- Queries on data cubes with examples
- Road map

INMOBI™

Queries on Data cubes

```
CUBE SELECT [DISTINCT] select_expr, select_expr, ...
FROM cube_table_reference
WHERE [where_condition AND] TIME_RANGE_IN(colName , from, to)
[GROUP BY col_list]
[HAVING having_expr]
[ORDER BY colList]
[LIMIT number]
```

```
cube_table_reference:
cube_table_factor
| join_table
join_table:
cube_table_reference JOIN cube_table_factor [join_condition]
| cube_table_reference {LEFT|RIGHT|FULL} [OUTER] JOIN cube_table_reference [join_condition]
cube_table_factor:
cube_name [alias]
| ( cube_table_reference )
join_condition:
ON equality_expression ( AND equality_expression )*
equality_expression:
expression = expression
colOrder: ( ASC | DESC )
colList : colName colOrder? (',' colName colOrder?)*
```

INMOBI™

Querying features

- Resolve candidate dimension tables and the storage tables .
- Resolve the candidate fact tables and the storage tables for the queried time range.
- Automatically resolve joins using the relationships between cubes and dimension.
- Automatically add aggregate functions to measures.
- Add expression to group by clause, if projected; and project group by clause, if it is not.

INMOBI™

Example query

cube select name, stateid from citytable limit 100

- SELECT (citytable . name), (citytable . stateid) FROM c2_citytable citytable LIMIT 100
- SELECT (citytable . name), (citytable . stateid) FROM c1_citytable citytable WHERE (citytable.dt = 'latest') LIMIT 100

INMOBI™

Example query

cube select citytable.name, msr2 from testcube where timerange_in(dt, '2014-03-10-03', '2014-03-12-03')

- SELECT (citytable.name), sum((testcube.msr2)) FROM c2_testfact testcube INNER JOIN c1_citytable citytable ON ((testcube.cityid)= (citytable.id)) WHERE ((testcube.dt='2014-03-10-03') OR (testcube.dt='2014-03-10-04') OR (testcube.dt='2014-03-10-05') OR (testcube.dt='2014-03-10-06') OR (testcube.dt='2014-03-10-07') OR (testcube.dt='2014-03-10-08') OR (testcube.dt='2014-03-10-09') OR (testcube.dt='2014-03-10-10') OR (testcube.dt='2014-03-10-11') OR (testcube.dt='2014-03-10-12') OR (testcube.dt='2014-03-10-13') OR (testcube.dt='2014-03-10-14') OR (testcube.dt='2014-03-10-15') OR (testcube.dt='2014-03-10-16') OR (testcube.dt='2014-03-10-17') OR (testcube.dt='2014-03-10-18') OR (testcube.dt='2014-03-10-19') OR (testcube.dt='2014-03-10-20') OR (testcube.dt='2014-03-10-21') OR (testcube.dt='2014-03-10-22') OR (testcube.dt='2014-03-10-23') OR (testcube.dt='2014-03-11') OR (testcube.dt='2014-03-12-00') OR (testcube.dt='2014-03-12 -01') OR (testcube.dt='2014-03-12-02'))AND (citytable.dt = 'latest') GROUP BY(citytable.name)

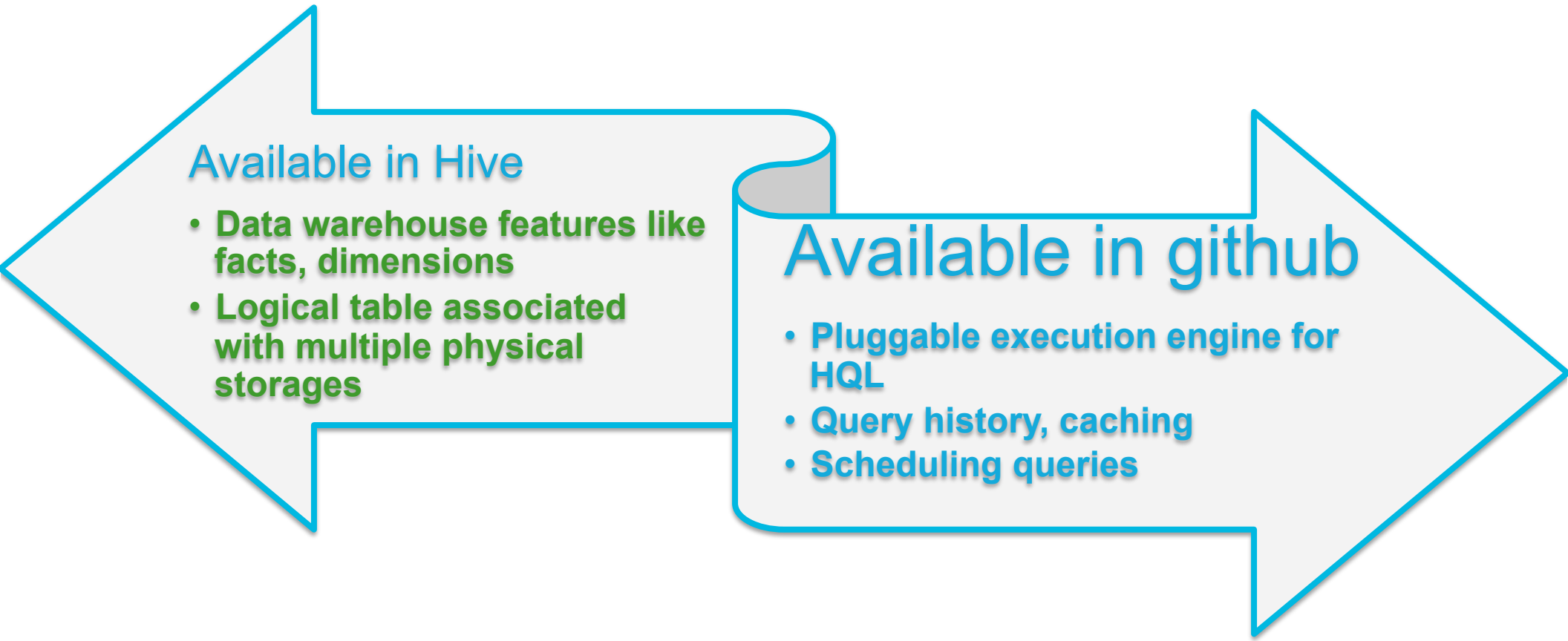
Data ware house statistics

Stats

- Number of queries - 700 to 900 per day
- Number of dimension tables - 125
- Number of fact tables – 24
- Number cubes – 15
- Size of the data
  - Total size – 136 TB
  - Dimension data – 400 MB compressed per hour
  - Raw data - 1.2 TB per day
  - Aggregated facts- 53GB per day

INMOBI™

What is available



INMOBI™

Pluggable execution engine

Implements an interface

- execute
- explain
- executeAsynchronously
- fetchResults
- Specify all storages it can support

INMOBI™

Cube query with multiple execution engines

Cube QL query

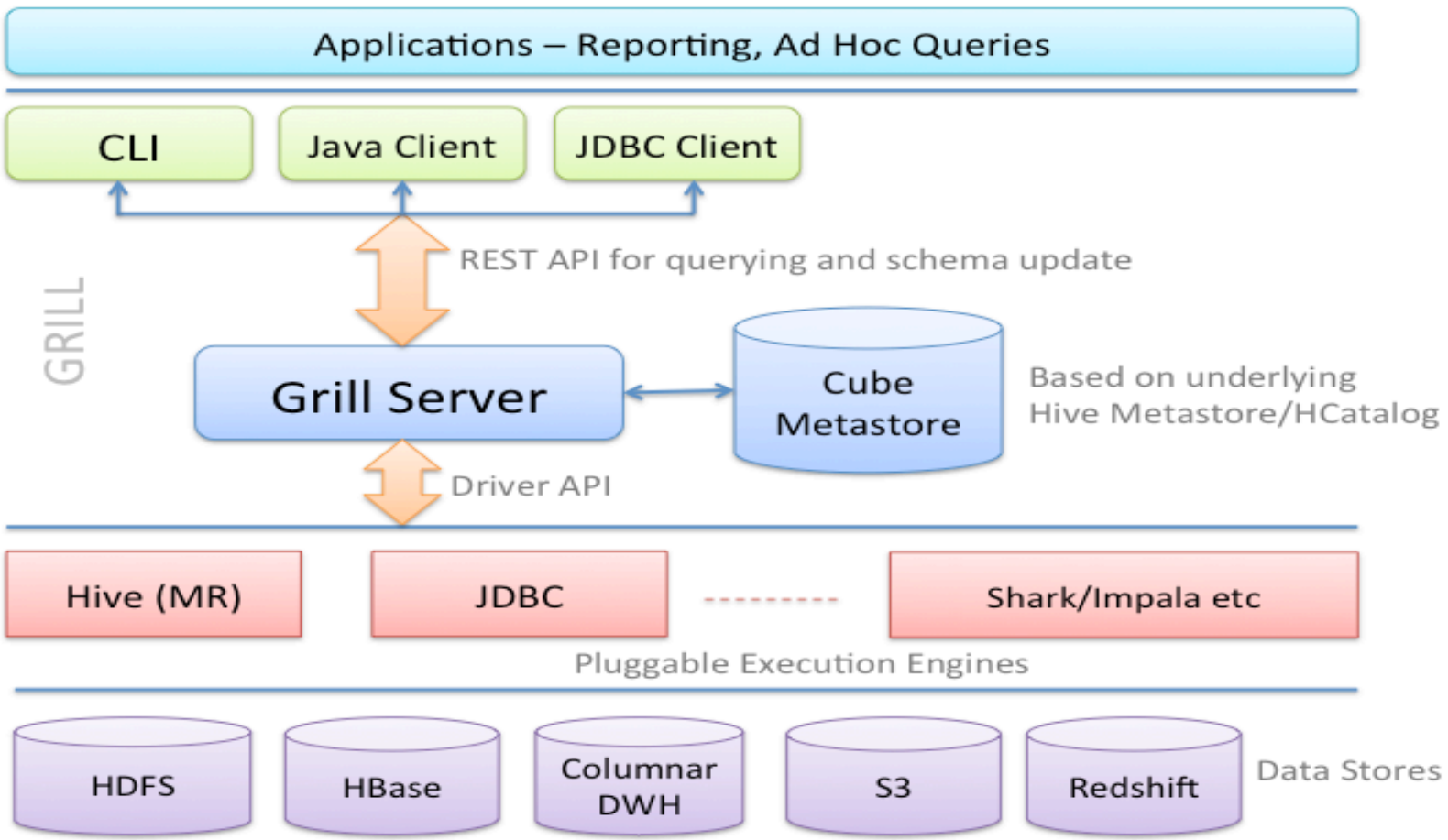
Rewrite query for available execution engines

Get cost of the rewritten query from each execution engine

Pick up execution engine with least cost and fire the query

INMOBI™

Future roadmap: Unified analytics



Thank you!

- amareshwari@apache.org
- jaideep.dhok@inmobi.com

INMOBI™