



A **very** short talk about Apache Kylin Business Intelligence meets Big Data

Fabian Wilckens
EMEA Solutions Architect



The challenge today ...



Very quickly: OLAP

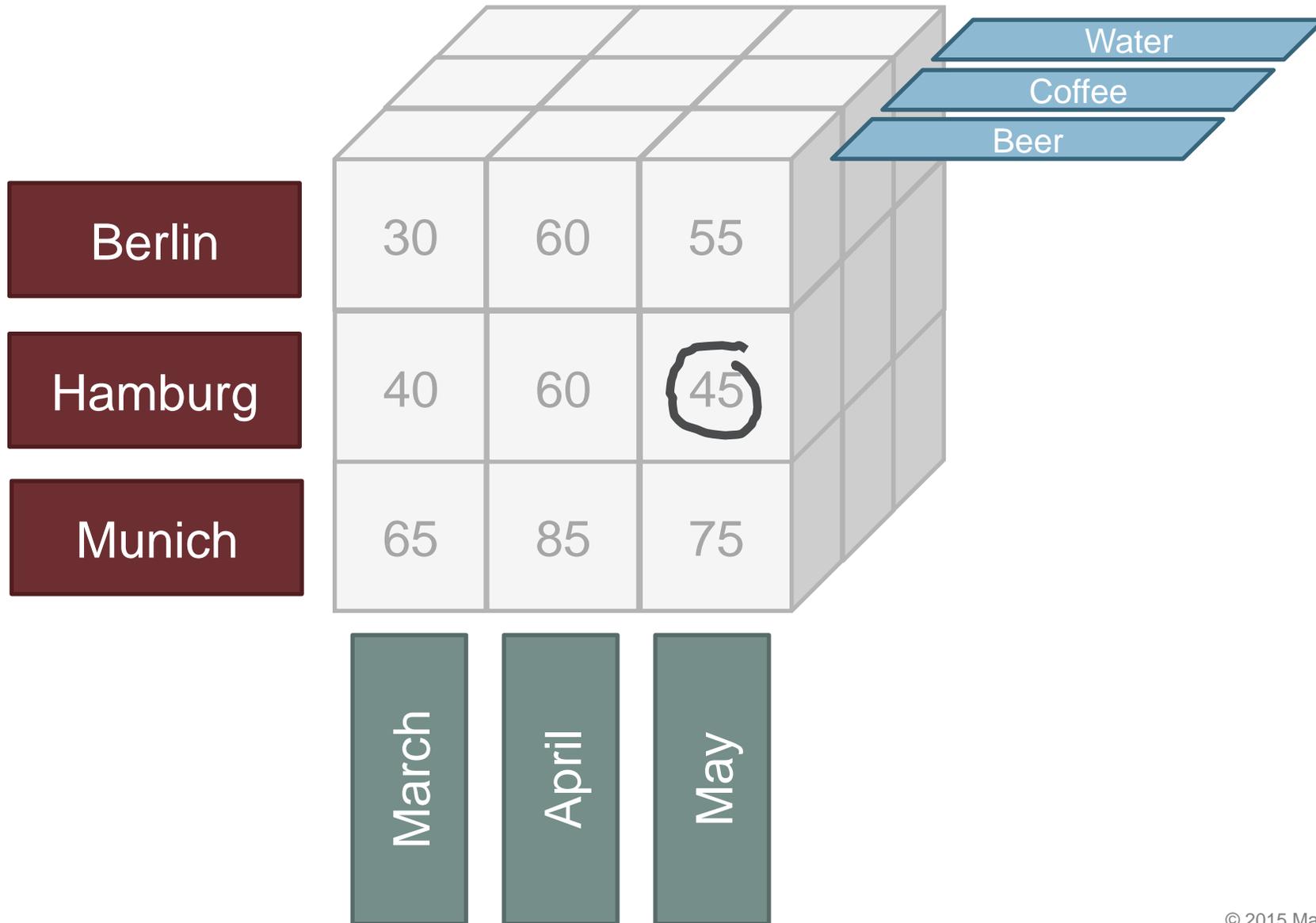
Online Analytical Processing



How many beers were ordered in Germany on a yearly basis?

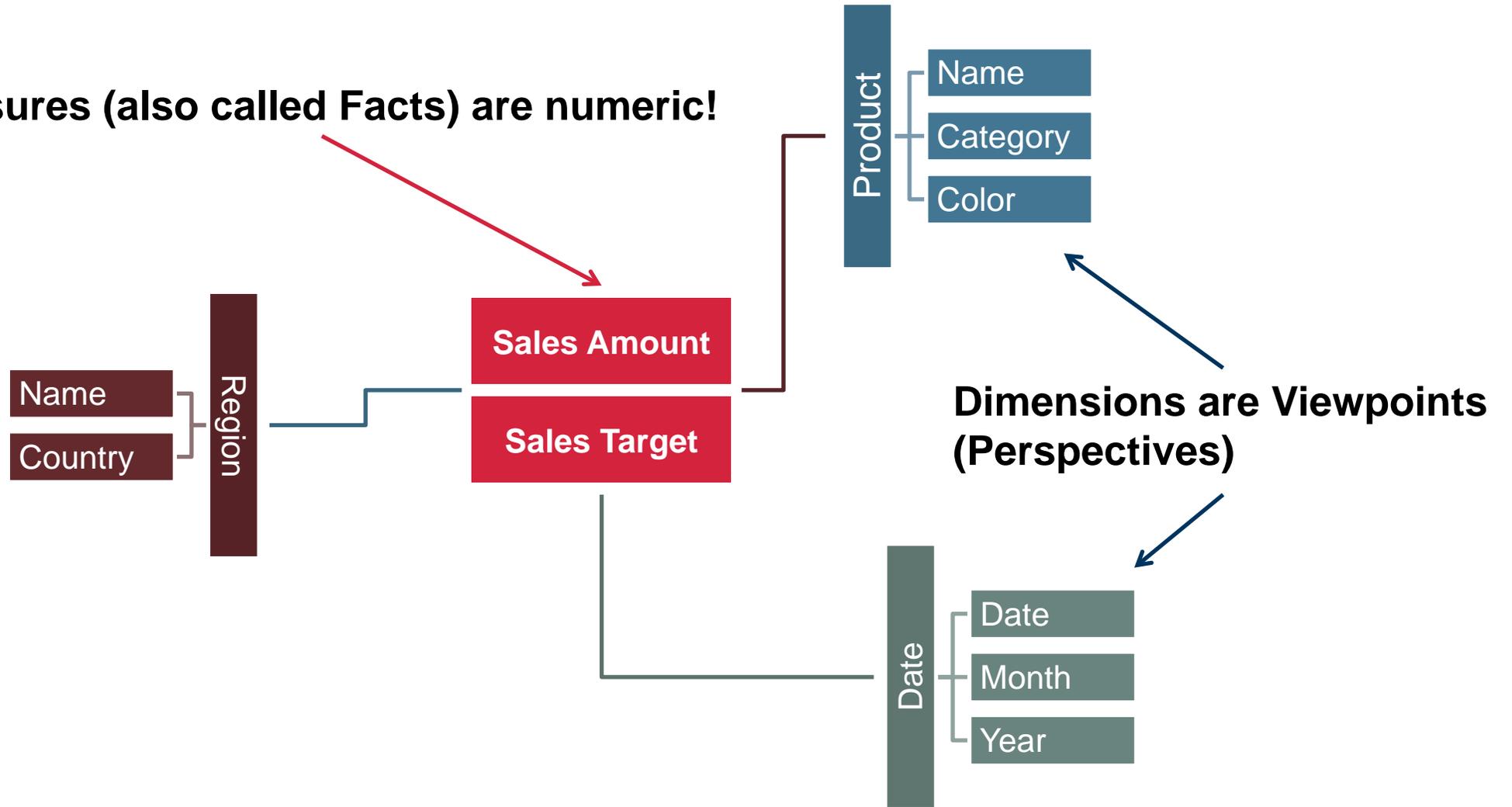
How many beers were ordered in Germany on a yearly basis, broken down by months?

OLAP Cubes

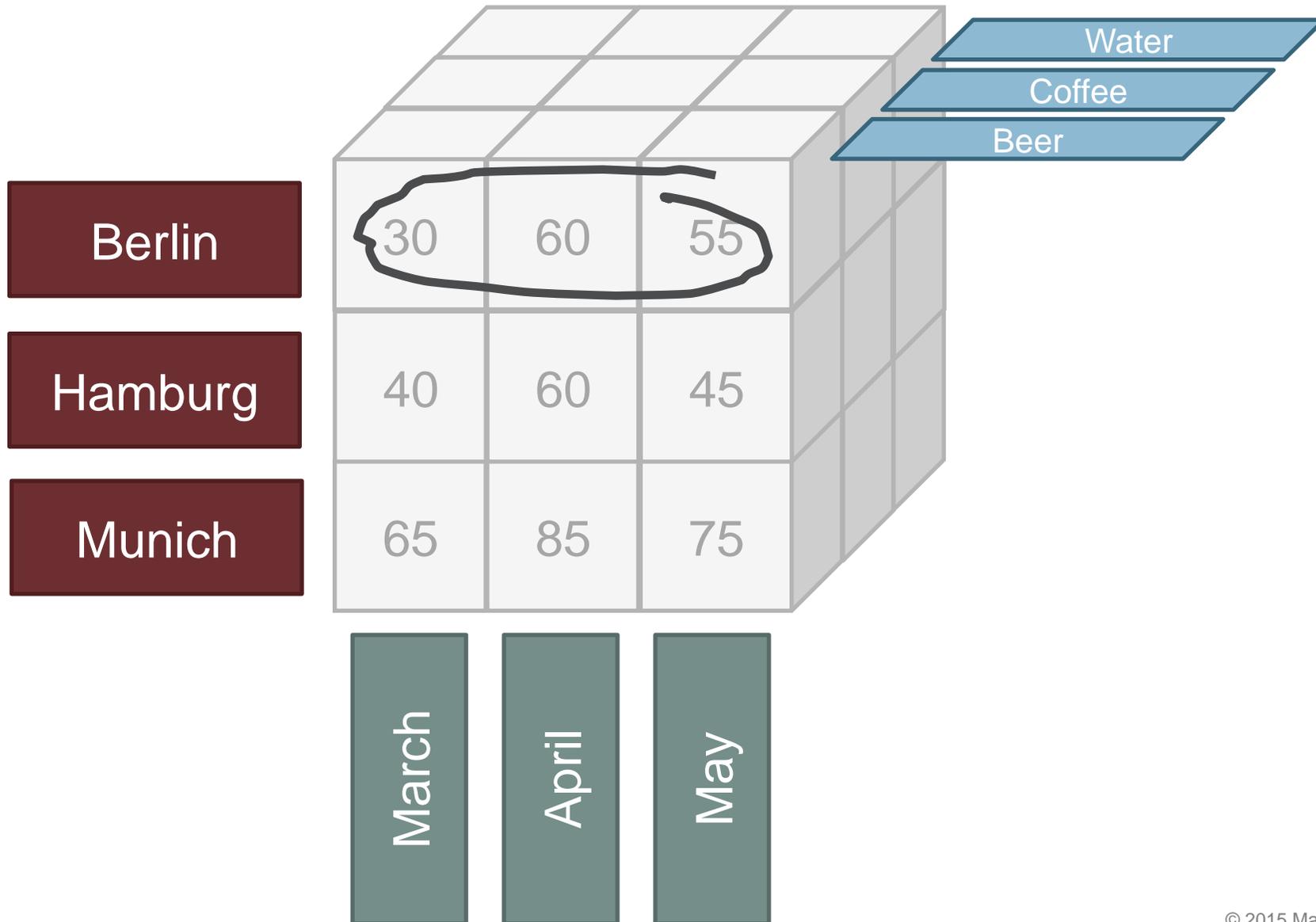


Cubes: Measures and Dimensions

Measures (also called Facts) are numeric!

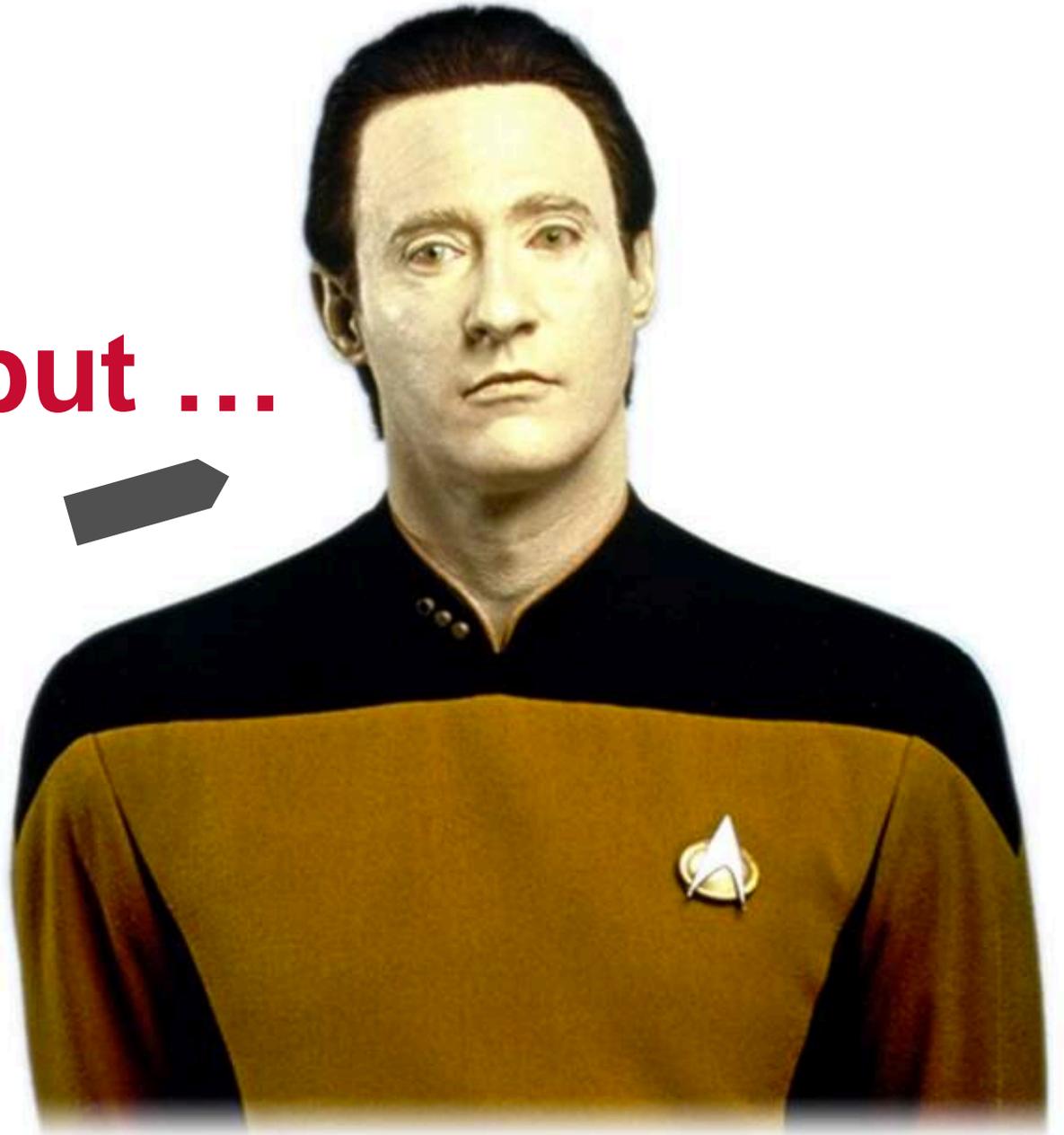


OLAP Cubes



That's all great but ...

How about **Big Data** 



What is Kylin?



kylin / 'ki:'lin / 麒麟

--n. (in Chinese art) a mythical animal of composite form

Extreme OLAP Engine for Big Data

Kylin is an open source Distributed Analytics Engine from (originally from eBay) that provides SQL interface and multi-dimensional analysis (OLAP) on Hadoop for extremely large datasets

- Open Sourced on Oct 1st, 2014
- Accepted into incubation November, 2014
- Preparing for first Apache release

Goals

- Sub-second query latency on billions of rows
 - ANSI SQL for both analysts and engineers
 - Full OLAP capability to offer advanced functionality
 - Seamless Integration with BI Tools
-
- Support for high cardinality and dimensionality
 - High concurrency – thousands of end users
 - Distributed and scale out architecture for large data volume

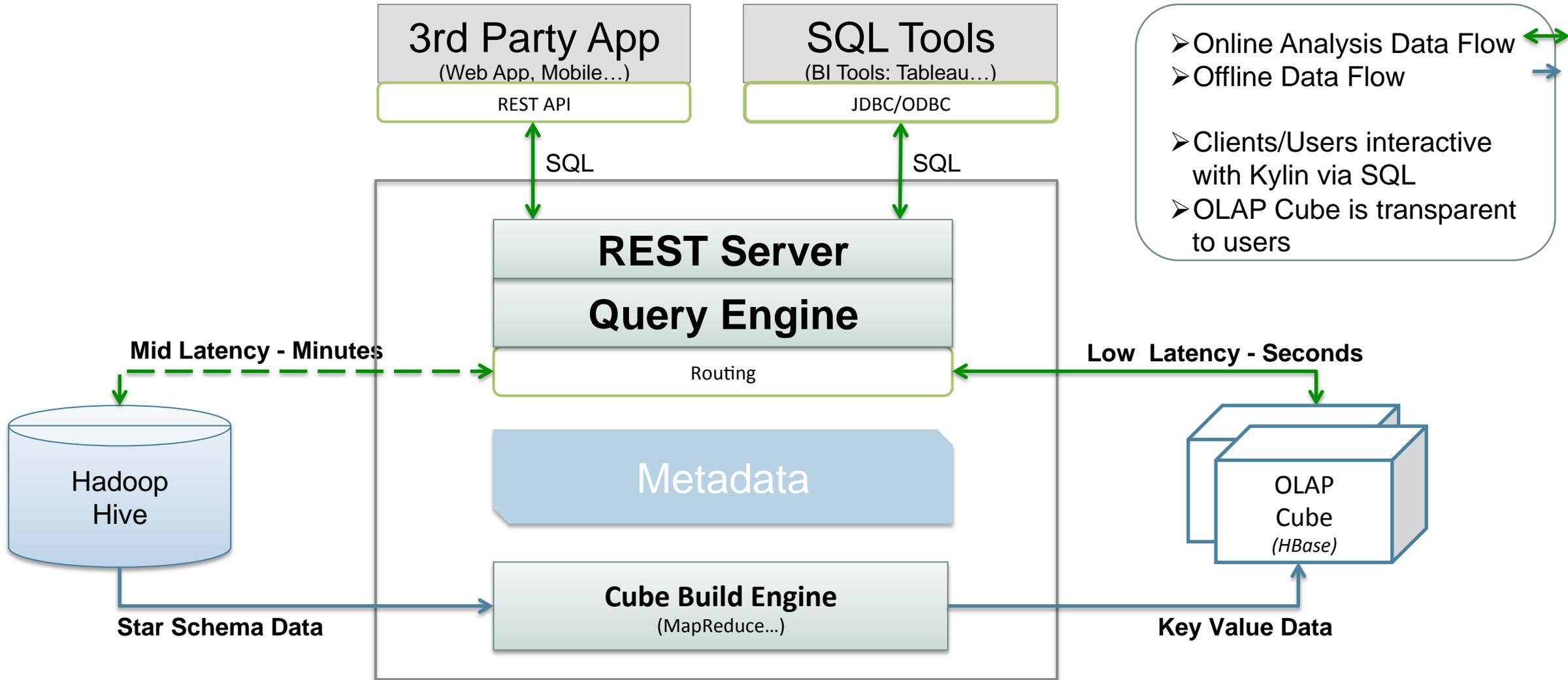


Kylin Depends on Hadoop Eco-system

- Hive
 - Input source, pre-join star schema during cube building
- MapReduce
 - Aggregate metrics during cube building
- HDFS
 - Store intermediate files during cube building
- HBase
 - Store and query data cubes
- Calcite
 - SQL parsing, code generation, optimization



Kylin Architecture Overview



Kylin Highlights

- ***Extremely Fast OLAP Engine at Scale***

Kylin is designed to reduce query latency on Hadoop for 10+ billions of rows of data to seconds

- ***ANSI SQL Interface on Hadoop***

Kylin offers ANSI SQL on Hadoop and supports most ANSI SQL query functions

- ***Seamless Integration with BI Tools***

Kylin currently offers integration capability with BI Tools like Tableau.

- ***Interactive Query Capability***

Users can interact with Hadoop data via Kylin at sub-second latency

- ***MOLAP Cube***

User can define a data model and pre-build in Kylin with more than 10+ billions of raw data records



More Highlights

- Compression and Encoding Support
- Incremental Refresh of Cubes
- Approximate Query Capability for distinct Count (HyperLogLog)
- Leverage HBase Coprocessor for query latency
- Job Management and Monitoring
- Easy Web interface to manage, build, monitor and query cubes
- Security capability to set ACL at Cube/Project Level
- Support LDAP Integration



Cube Designer

Kylin Query Cubes Jobs Tables Admin Help Welcome, AD

Source Tables Cube Designer

test_kylin_cube_with_slr_empty	READY	6.52 MB	10,000	2014-11-11 13:59:15	Action	Action
--------------------------------	-------	---------	--------	---------------------	--------	--------

Grid Visualization SQL JSON Access Notification HBase

```
graph LR; F[TEST_KYLIN_FACT] --- I1[inner join] --- D1[TEST_CAL_DT]; F --- I2[inner join] --- D2[TEST_CATEGORY_GROUPINGS]; F --- I3[inner join] --- D3[TEST_SITES]; F --- I4[inner join] --- D4[TEST_SELLER_TYPE_DIM];
```

Setting 6 Advanced Setting 7 Overview

Tips

1. Please indicate which type for refresh model
2. Leave as default if this cube always need full build
3. Please indicate partition column of Fact Table in Hive
4. Partition column accept expression like: concat(year, '-', month, '-', day)
5. Please indicate start date to just pull certain data from source

← Prev Next →

Job Management

The image displays a collage of screenshots related to job management in a data warehouse environment, specifically Kylin and Hadoop.

Top Left Screenshot (Kylin UI): Shows a navigation bar with 'Jobs', 'Tables', and 'Admin'. Below it, a 'CUBE BUILD CONFIRM' dialog box is visible, showing 'PARTITION DATE COLUMN' as 'test_kylin_fact.cal_dt' and 'START DATE' as 'Thursday, January 1, 1970'.

Middle Left Screenshot (Kylin Jobs List): A table listing jobs with columns for 'Job Name', 'Duration', and 'Actions'. Two jobs are listed:

Job Name	Duration	Actions
test_kylin_cube_with_str_empty - 19700101000000_19710101000000 - BUILD - PST 2014-11-05 03:01:20	0.00 mins	Action [Refresh]
test_kylin_cube_without_str_empty - FULL_BUILD - BUILD - PST 2014-11-05 01:28:57	24.07 mins	Action [Refresh]

Middle Right Screenshot (Job Detail Information): Shows 'Detail Information' for a job named 'test_kylin_cube_without_str_empty - FULL_BUILD - BUILD - PST 2014-11-05 01:28:57'.

Bottom Left Screenshot (Job Status): A black box containing the following text:

```
SequenceID: 0
Status: FINISHED
Duration: 5.44 mins
Waiting: 0 seconds
Start At: 2014-11-05 17:29:08
End At: 2014-11-05 17:34:34
Data Size: 446.61 KB
MR Job:
job_1415168056392_0001
```

Bottom Middle Screenshot (Output Log): Shows the output of a job, including 'Logging initialized using configuration in file:', 'OK', 'Time taken: 7.782 seconds', 'Query ID = root_20141105012929_23a1dace-ba6b-4af', 'Total jobs = 1', and '2014-11-05 01:31:18 Starting to launch local'.

Bottom Right Screenshot (Hadoop MapReduce Job Details): Shows the details of a Hadoop MapReduce job 'MapReduce Job job_1415168056392_0001'. The job is in a 'SUCCEEDED' state. The output table shows the following statistics:

Task Type	1	Total	1	Complete
Map	1	1	0	0
Reduce	0	0	0	0
Attempt Type	Failed	Killed	Successful	
Map	0	0	1	
Reduce	0	0	0	

Query and Visualization

Results (1215)

Visualization Export

Drag a column header here and drop it to group by that column.

LSTG_FORMAT...	WEEK_BEG_DT	META_CATEG...	PRICE
FP-GTC	2013-01-01	Sports MeCard...	93.268450871...
Auction	2013-01-01	Business & Ind...	55.635632631...
Others	2013-01-01	Sporting Goods	82.533676494...
Auction	2013-01-01	Real Estate	20.347844733...
Auction	2013-01-01	Toys & Hobbies	4.0906749147...

Results

Status: Success

Project: onlyinner

Cubes: test_kylin_cube_with_slr_empty

Start Time: 2014-11-12 17:33:48 Duration: 3.10s

Query String

Results (1215)

Graph Type: Pie Chart

Dimensions: LSTG_FORMAT...

Metrics: PRICE

Category	Color
ABIN	Blue
Auction	Light Blue
FP-GTC	Orange
FP-non GTC	Light Orange
Others	Green

Results (1215)

Drag a column header here and drop it to group by that column.

LSTG_FORMAT...	WEEK_BEG_DT	META_CATEG...
FP-GTC	2013-01-01	Spo
Auction	2013-01-01	Bus

Results

Status: Success

Project: onlyinner

Cubes: test_kylin_cube_with_slr_empty

Start Time: 2014-11-12 17:33:48 Duration: 3.10s

Query String

Results (1215)

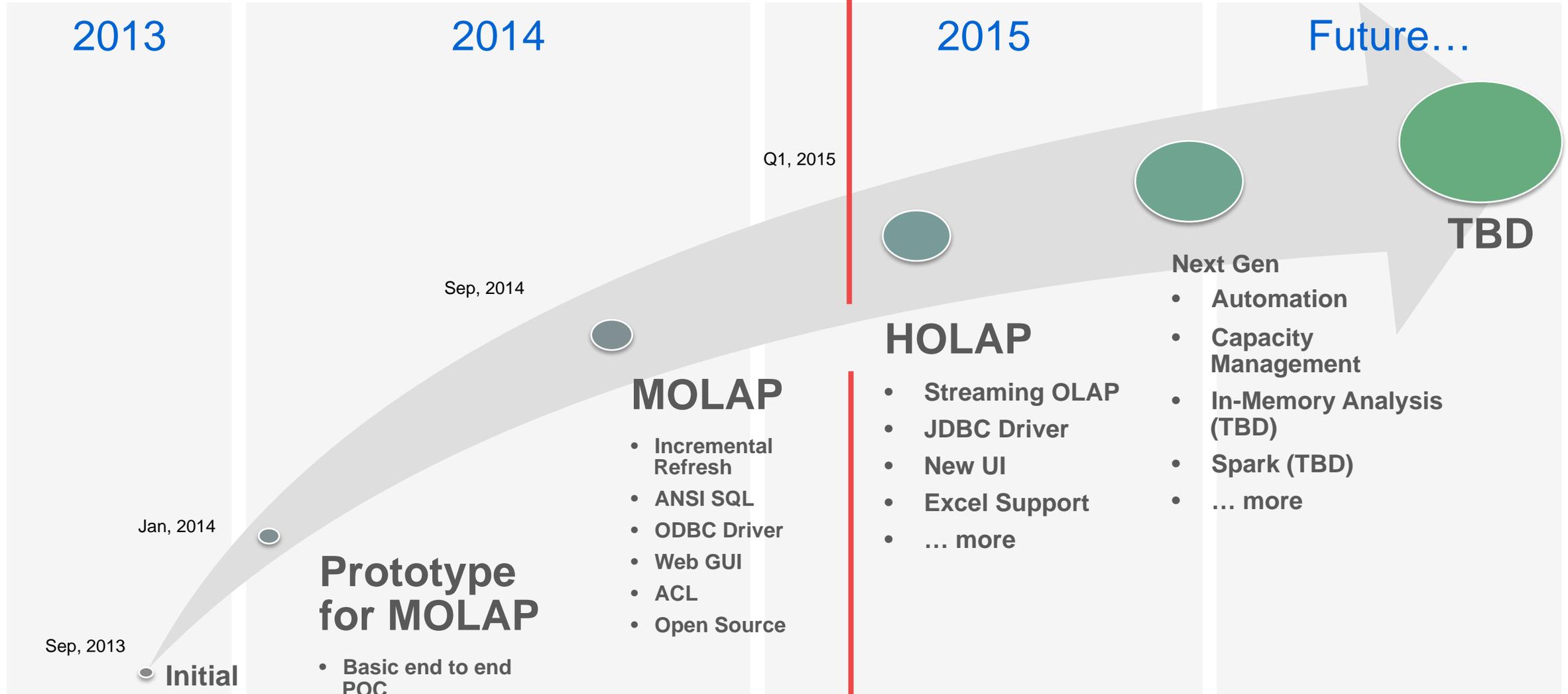
Graph Type: Line Chart

Dimensions: WEEK_BEG_DT

Metrics: PRICE

Date	Price
2013-01-01	~4000
2013-02-05	~4500
2013-02-28	~5000
2013-03-23	~5500
2013-04-15	~5000
2013-05-08	~5500
2013-05-26	~6000

Kylin History and Roadmap



Open Source

- Kylin Site:
 - <http://kylin.io>
- Twitter:
 - [@ApacheKylin](https://twitter.com/ApacheKylin)
- Source Code Repo:
 - <https://github.com/KylinOLAP>
- Google Group:
 - [Kylin OLAP](#)
- 微信
 - ApacheKylin

The screenshot shows the Apache Kylin website. At the top, there is a navigation bar with links for HOME, DOCS, COMMUNITY, GITHUB, and ABOUT. Below this is a section titled "APACHE KYLIN OVERVIEW". The text states that Kylin was accepted as an Apache Incubator Project on Nov 25, 2014, and is an open source Distributed Analytics Engine from eBay Inc. that provides a SQL interface and multi-dimensional analysis (OLAP) on Hadoop for large datasets.

A diagram illustrates the architecture. On the left, "Hadoop Hive" provides "Star Schema Data" to the "Cube Build Engine". The "Cube Build Engine" outputs "Key Value Data" to "HBase as Storage". The "OLAP Cube" is then accessed by a "REST Server" (via "REST API") and a "Query Engine" (via "JDBC/ODBC"). The "Query Engine" and "REST Server" both send "SQL" requests to the "Routing" layer. The "Routing" layer then sends "Low Latency-Seconds" data to the "OLAP Cube". A box on the right lists features: Online Analysis Data Flow, Offline Data Flow, Only SQL for End User, and OLAP Cube is transparent to users. A red banner with white text "Fork me on GitHub" is overlaid diagonally across the top right of the screenshot.





Fabian Wilckens
fwilckens@mapr.com

Free on-demand
Hadoop training leading to certification
Start becoming an expert now
mapr.com/training

