eBay Real-time OLAP Engine Built on Apache Kylin

May 2017

What is Apache Kylin

ebay CHINA CENTER OF EXCELLENCE

kylin 麒麟

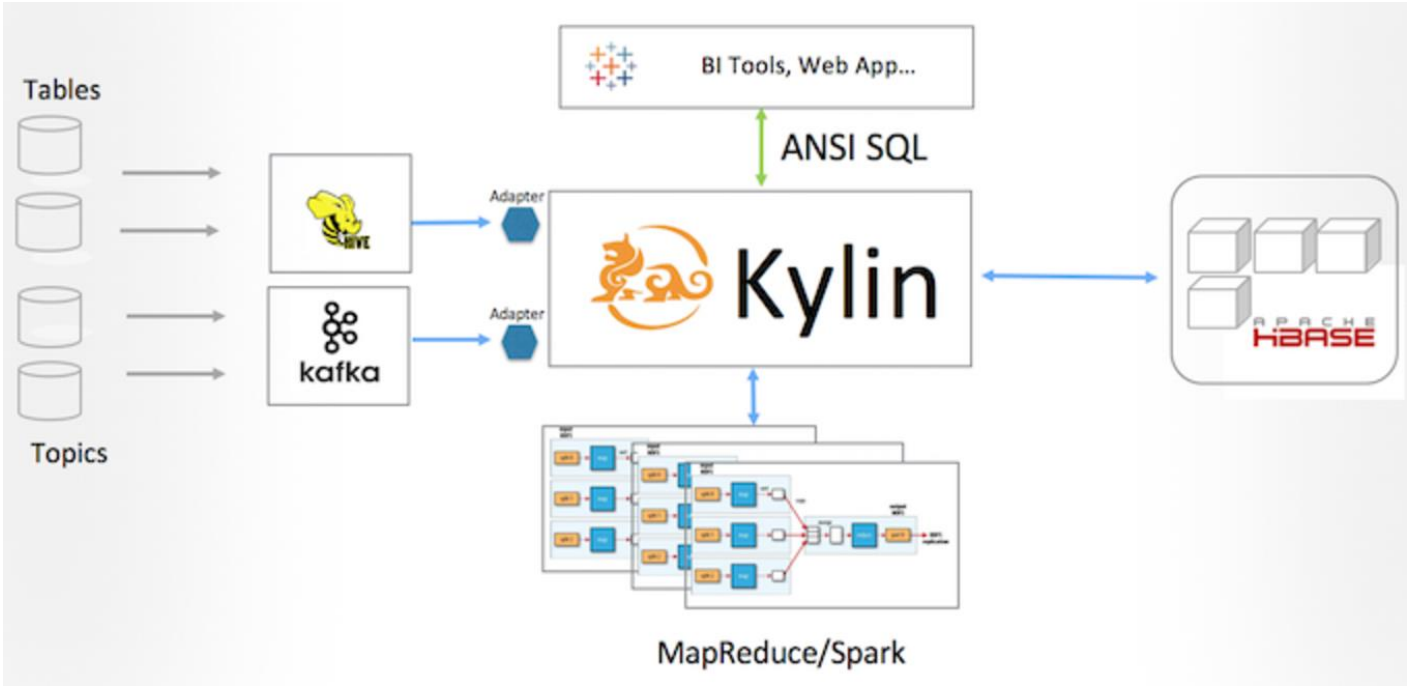
--n. (in Chinese art) a mythical animal of composite form

Extreme OLAP Engine for Big Data

Apache Kylin is an open source Distributed Analytics Engine designed to provide SQL interface and multi-dimensional analysis (OLAP) on Hadoop supporting extremely large datasets, initiated and continuously contributed by eBay Inc.

Kylin fills an important gap – OLAP on Hadoop

NRT Streaming in Apache Kylin 1.6



Current Design Limitation

ebay
CHINA CENTER
OF EXCELLENCE

- Minutes level build latency
- Create many tiny HBase tables
- Highly depends on Kafka
- Hard to combine Hive and Kafka together

Real-time OLAP

Milliseconds Query Latency + Milliseconds Data Preparation Delay

The Challenge

ebay CHINA CENTER OF EXCELLENCE

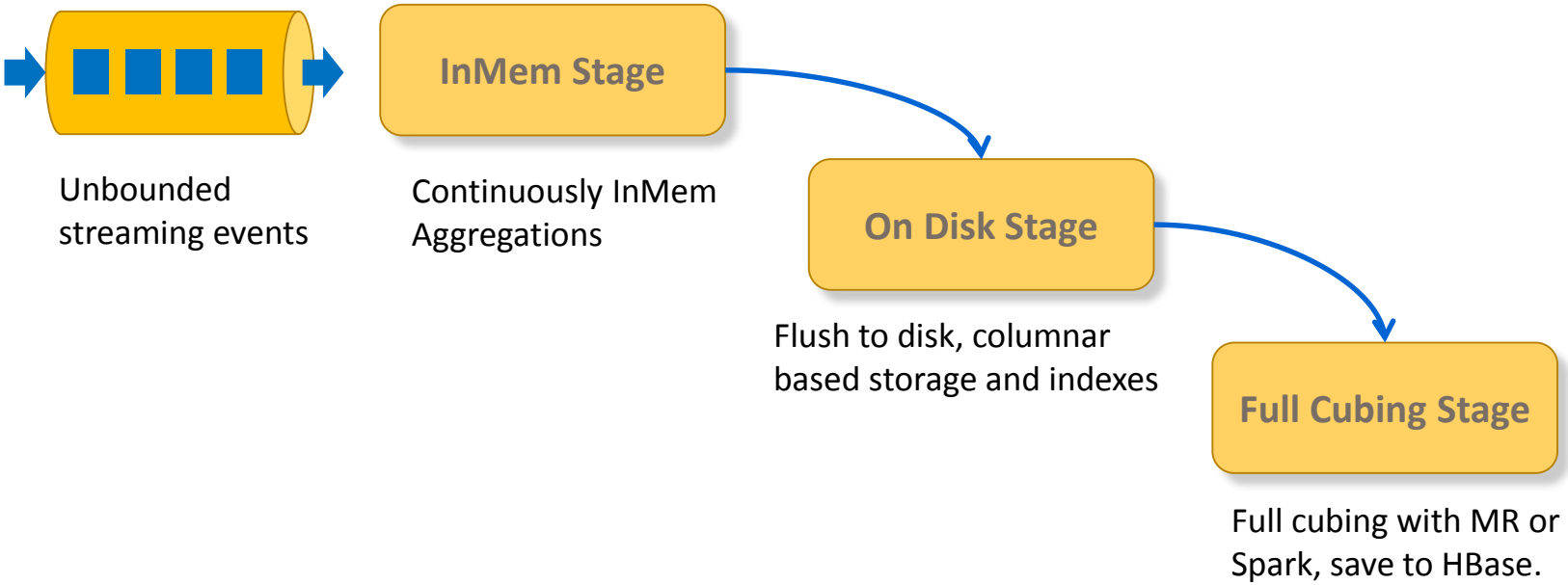
Remember Kylin is about pre-aggregations, it successfully reduced the query latencies by spending more time and resources on data preparation.

But for real-time OLAP, cubing time and data visibility latency is critical. Real-time applications are more sensitive to resource usage also.

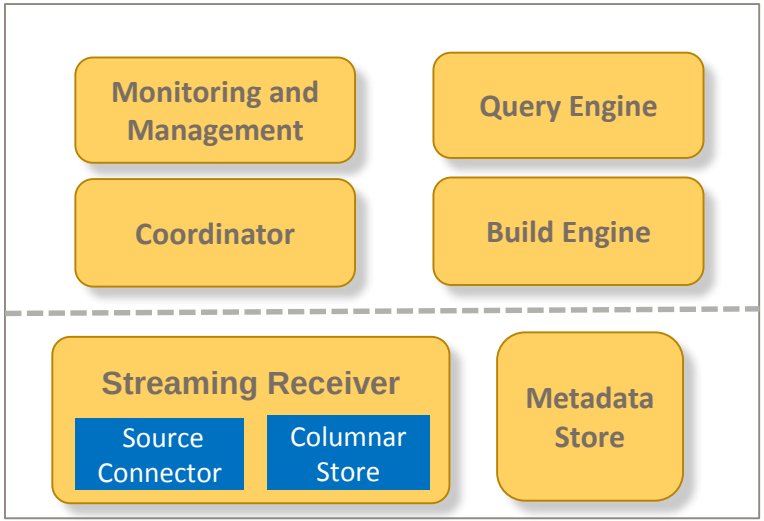
Can we **pre-build the cubes in real-time** in a more **cost effective** manner? This is hard but still doable.

The New Solution

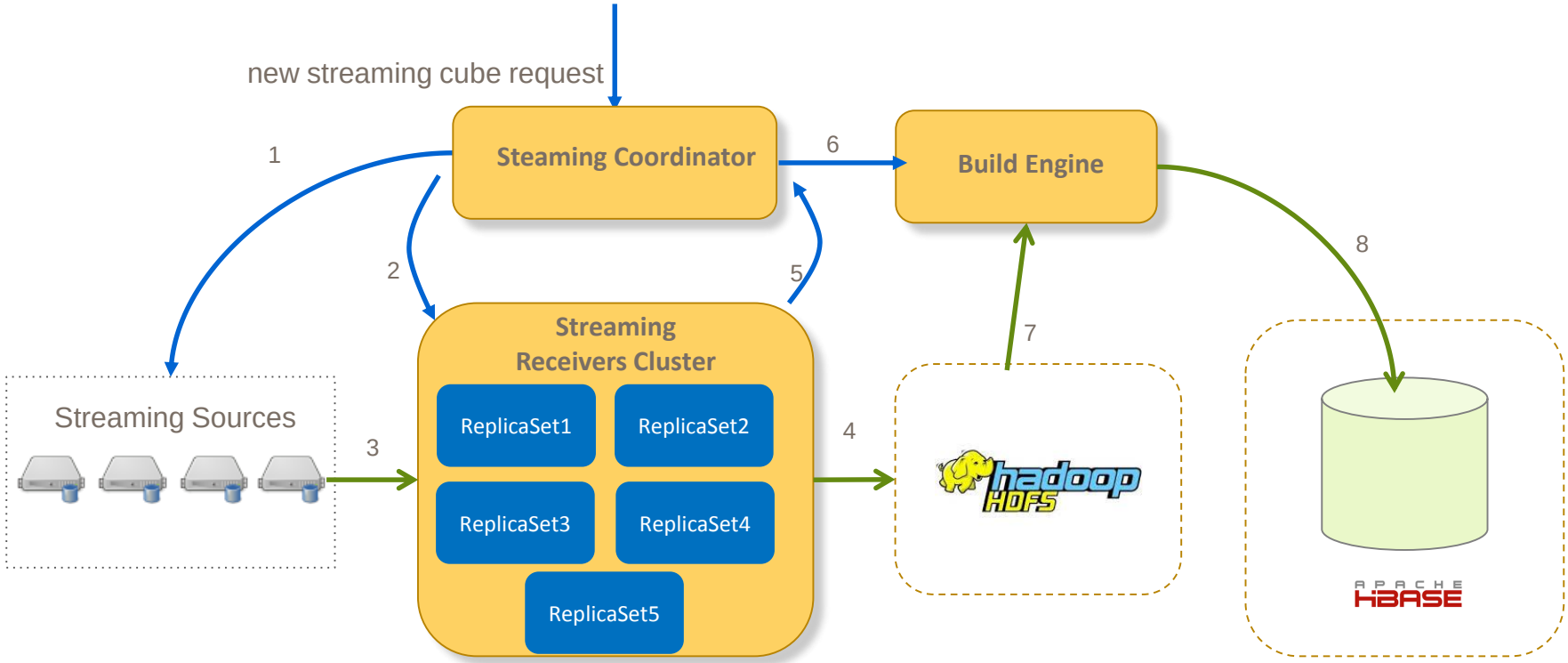
We divide the unbounded incoming streaming data into 3 stages, the data come into different stages are all queryable.



New Streaming Components



How Streaming Cube Engine Works

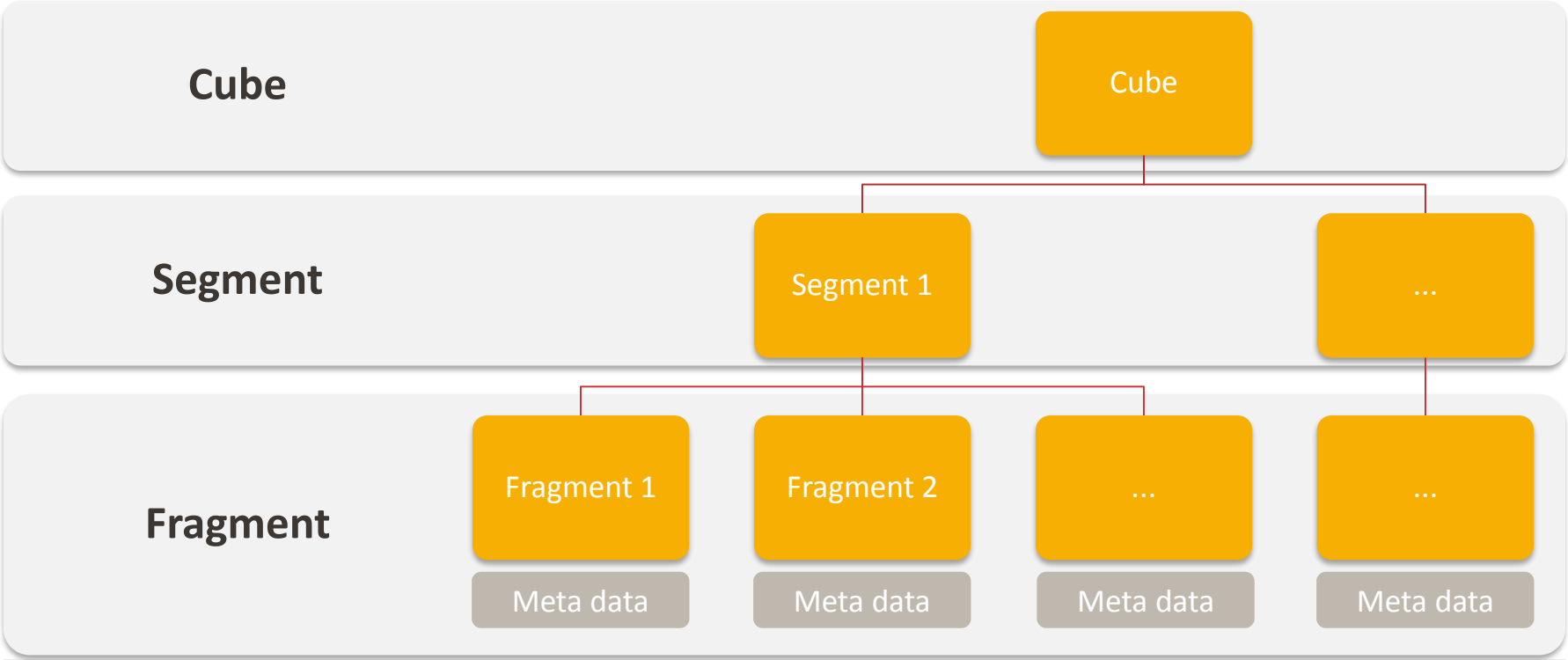


Core concepts

ebay
CHINA CENTER
OF EXCELLENCE

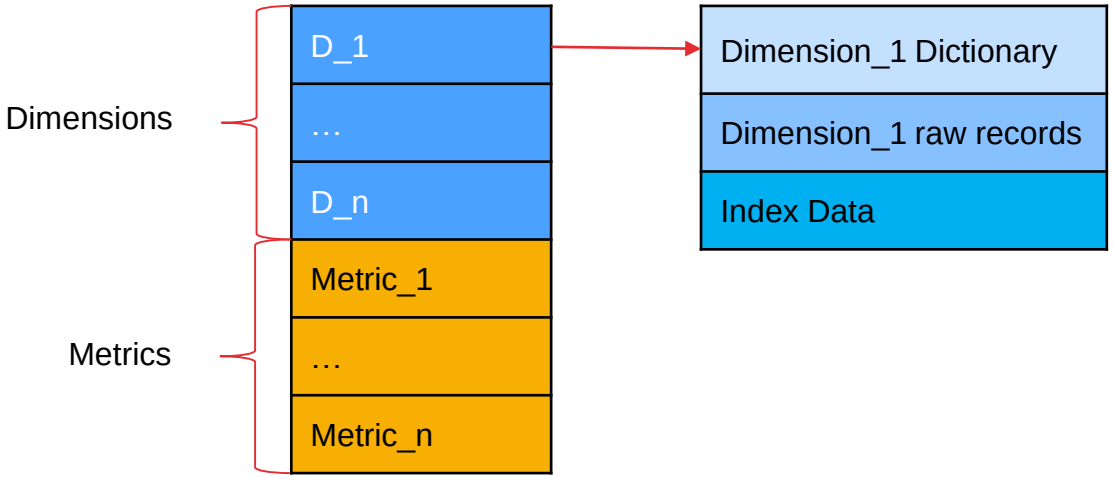
- Hierarchy Data Storage
- On Disk Columnar Store
- Segment Window
- Active Segments
- Immutable Segments
- Replica Set
- Check Point

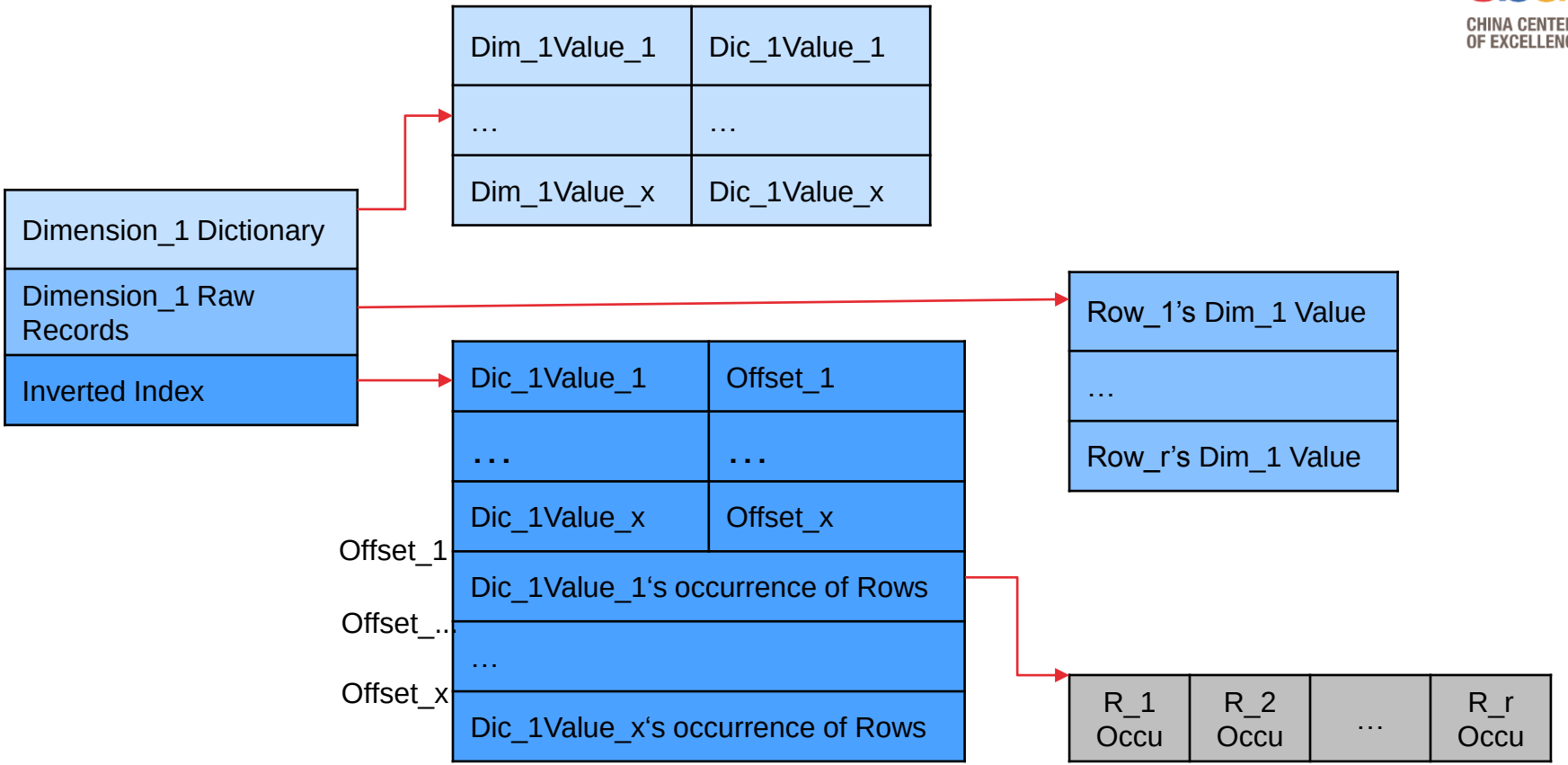
Hierarchy Data Storage



Column Based Fragment File Design

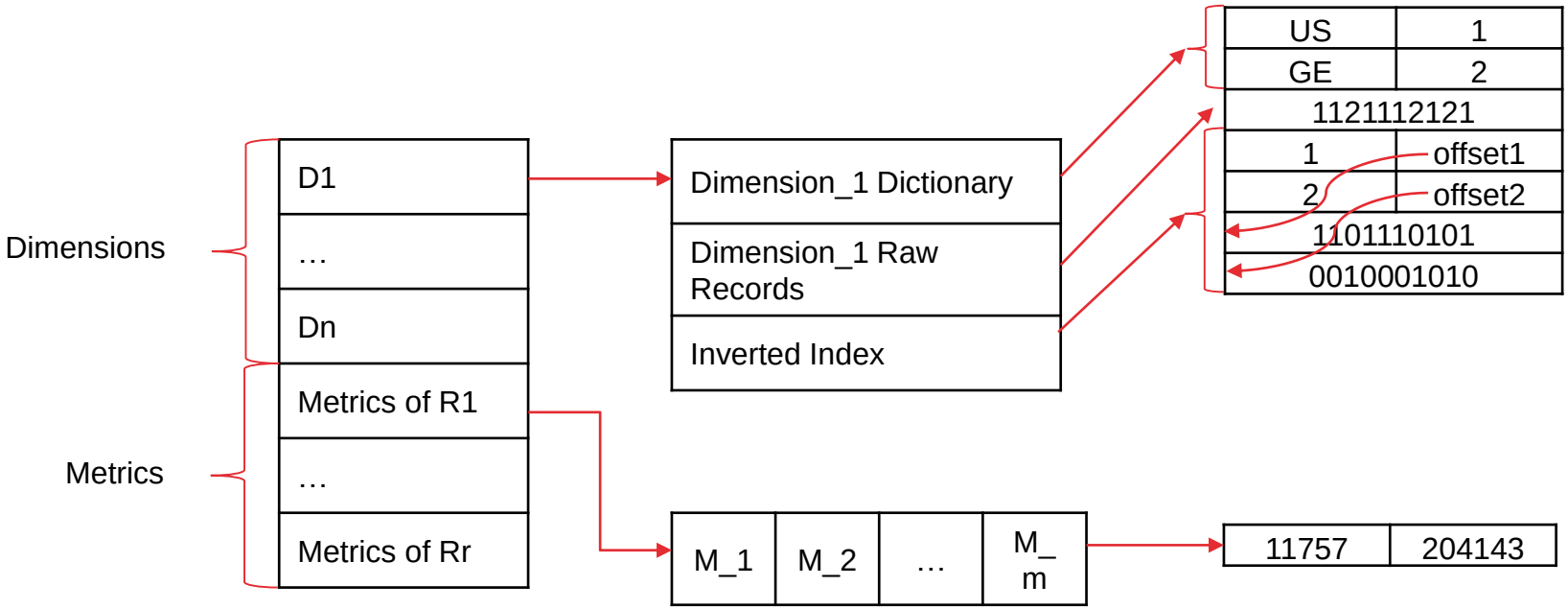
Dimensions : D1,D2…Dn.
Metrics : M1,M2...Mm.
RawRecords : R1,R2...Rr.




Dim_1Value_1
Dic_1Value_1
…
…
Dim_1Value_x
Dic_1Value_x
Dimension_1 Dictionary
Dimension_1 Raw Records
Inverted Index
Row_1's Dim_1 Value
…
Row_r's Dim_1 Value
Dic_1Value_1
Offset_1
…
…
Dic_1Value_x
Offset_x
Offset_1
Dic_1Value_1's occurrence of Rows
Offset_...
…
Offset_x
Dic_1Value_x's occurrence of Rows
R_1 Occu
R_2 Occu
…
R_r Occu

Dimensions : Site, Category, Date
Metrics : SPS, PV.
RawRecords :

Site	Category	Date	SPS	PV
US	Collectibles	8/27/16	11757	204143
US	Collectibles	8/28/16	10949	192350
GE	Fashion	8/28/16	5338	186659
US	Collectibles	8/29/16	9609	186283
US	Fashion	8/27/16	15174	177123
US	Fashion	8/28/16	14612	172552
GE	Home & Garden	8/27/16	7336	171364
US	Fashion	8/29/16	13927	171235
GE	Home & Garden	8/28/16	8615	169846
US	Home & Garden	8/29/16	4560	166506

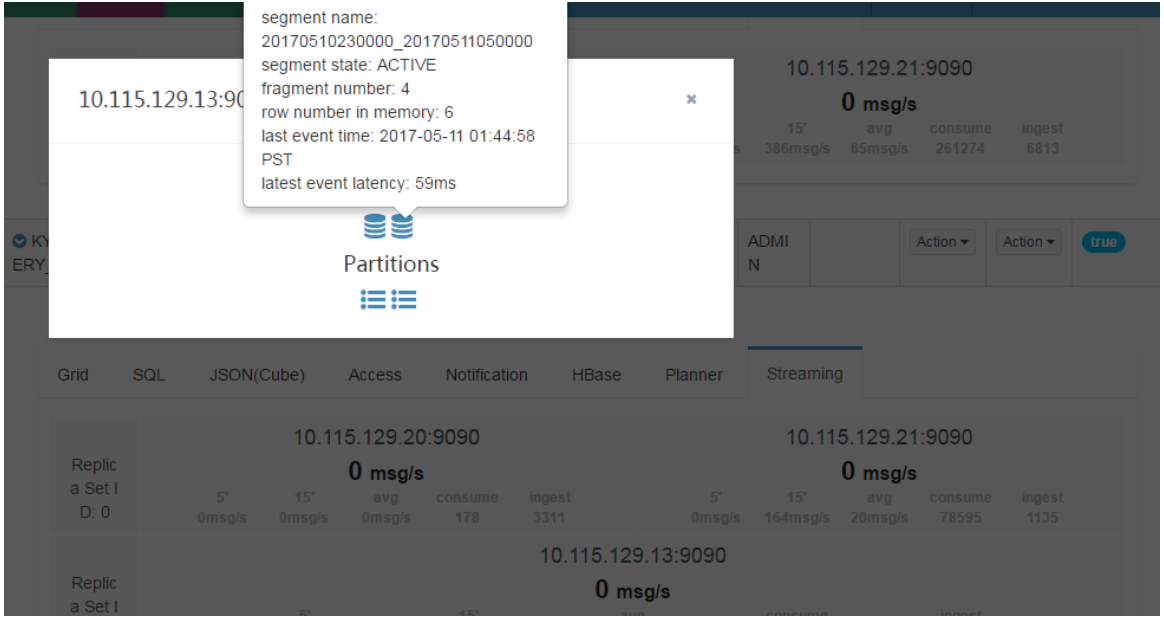

Dimensions
Metrics
D1
…
Dn
Metrics of R1
…
Metrics of Rr
Dimension_1 Dictionary
Dimension_1 Raw Records
Inverted Index
US
1
GE
2
1121112121
1
offset1
2
offset2
1101110101
0010001010
M_1
M_2
…
M_ m
11757
204143

Segment Window

Segments are divided by event time.

New created segments are first
In Active state.

After a preconfigured data retention period, no further events coming, The segment will be changed to Immutable state and then write to remote HDFS.



Long latency events can't fit into active Segments will write to long latency segment.

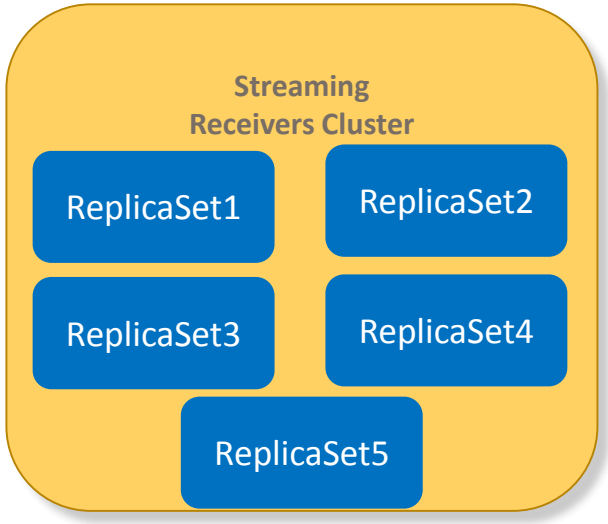
Replica Set

The concept of Replica Set is similar to many other distributed systems like Kafka, Mongo, Kubernetes, etc.

Replica Set ensures the replications for HA purpose.

Streaming receiver instances in the same replica set have the exact same local state.

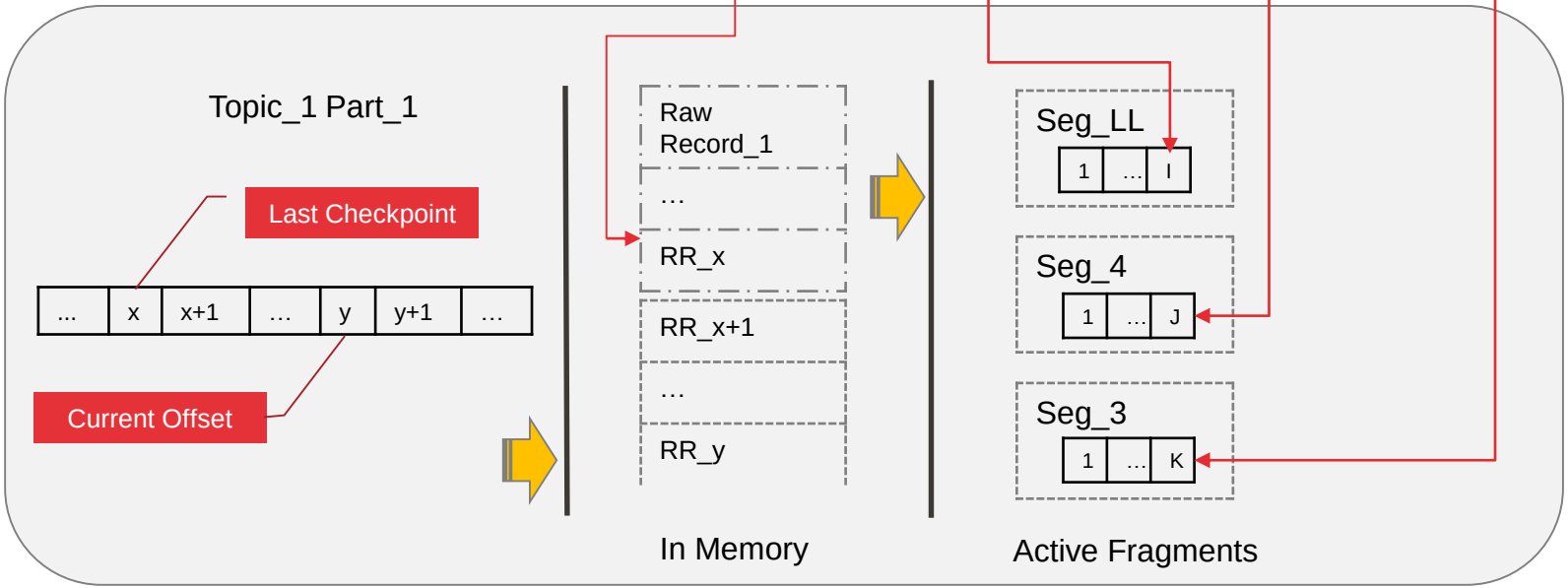
Streaming receive instances are pre-allocated to join the ReplicaSet groups.



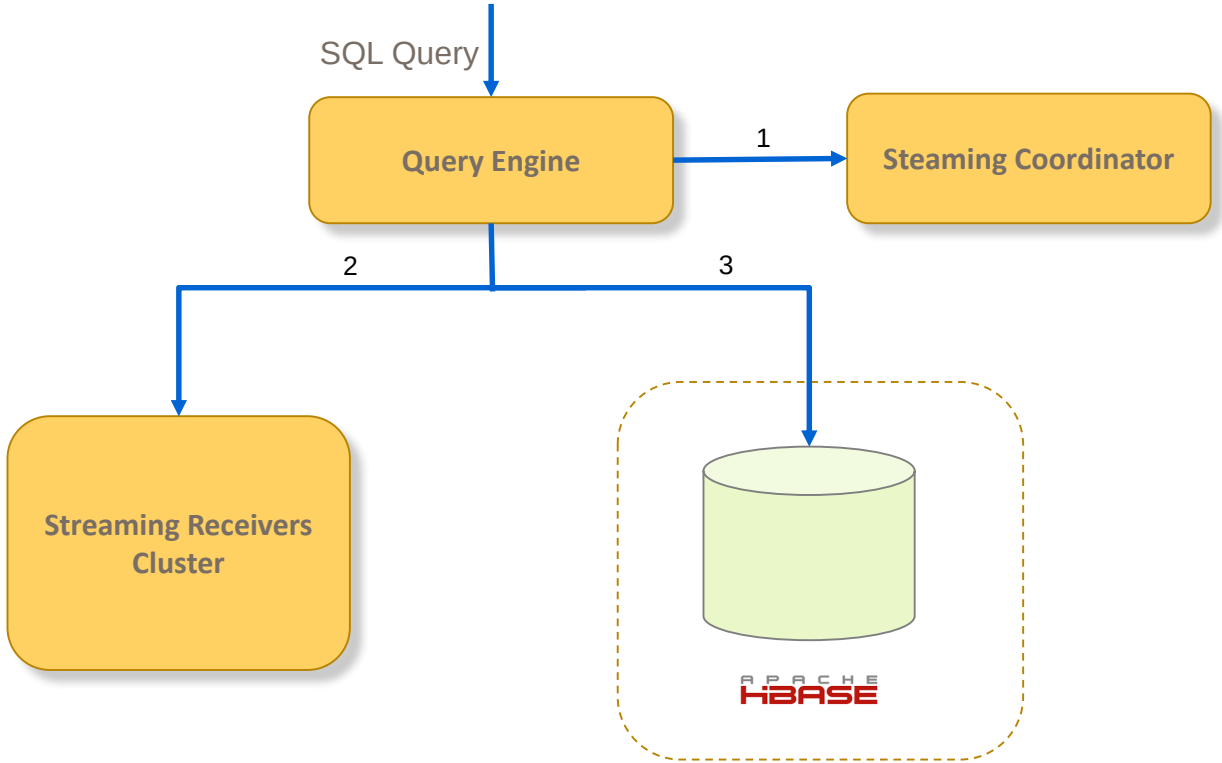
Check Point

Kafka

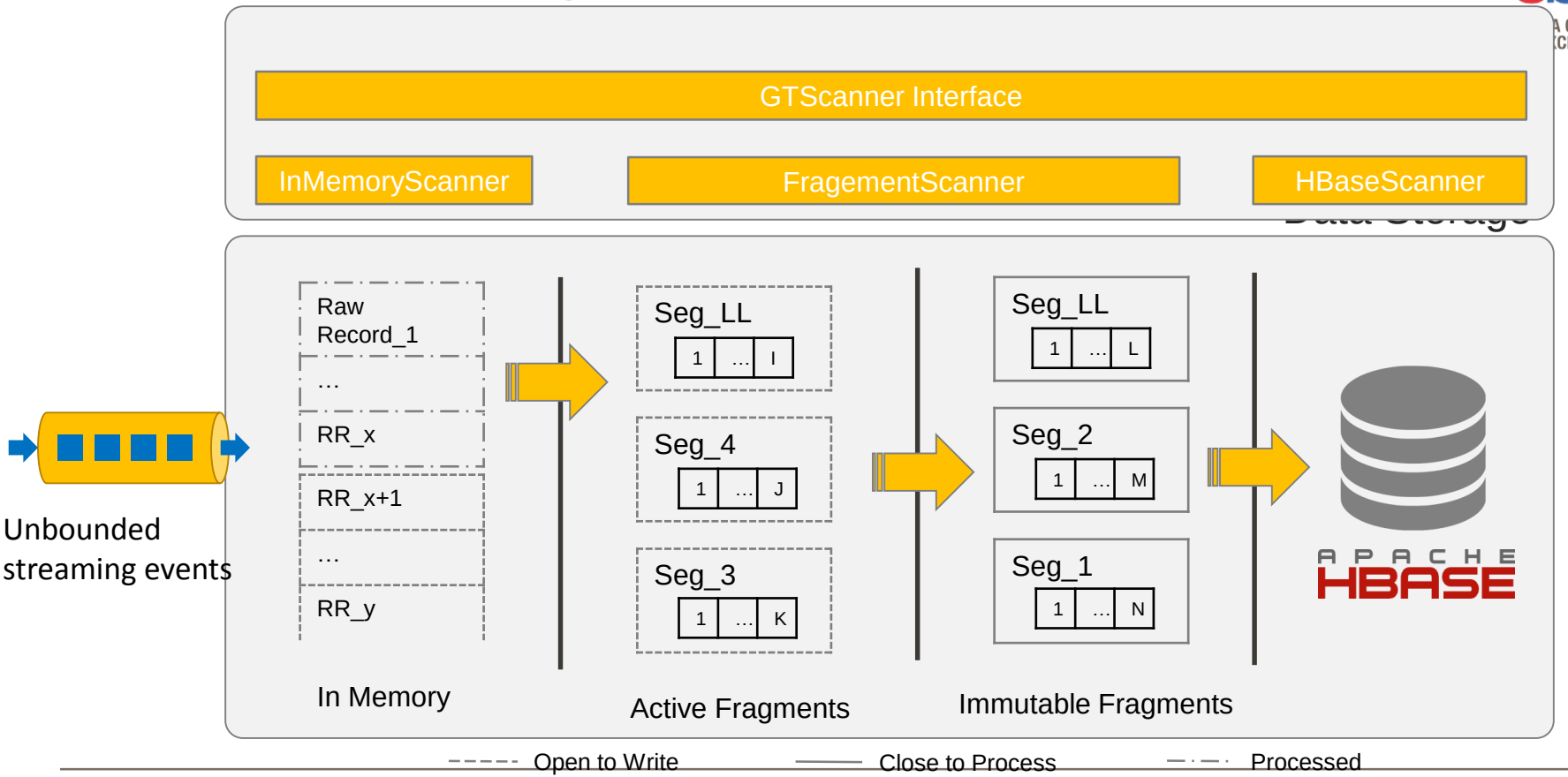
Date Time	Topic Name	Partition #	Offset	SeqID of Seg_LL	SeqID of Active Segments	
2016/10/01 12:00:00	Topic_1	1	x	I	Seg_4, J	Seg_3, K



Extended Query Engine



Extended Query Engine



Cubing Job

ebay CHINA CENTER OF EXCELLENCE

Coordinator submit MR jobs incrementally to build the full cube when all the streaming receiver instances convert their segment to Immutable for that window.

The MapReduce jobs will do all the dirty work like merge all fragment files together, build dictionaries, build cubes, generate Hfiles and bulkload to Hbase.

Incrementally merge jobs can be configured to merge hourly segments
to daily and then to weekly segments to avoid too many Hbase tables

Cubing Job

KYLIN_RHEOS_METRICS_QUERY_QA2 - 20170509170000_20170509230000 - STREAM - PDT 2017-05-10 23:23:43	KYLIN_RHEOS_METRICS_QUERY_QA2	100%	2017-05-10 23:30:21 PST	6.37 mins	Action	
KYLIN_RHEOS_METRICS_QUERY_QA2 - 20170509110000_20170509170000 - STREAM - PDT 2017-05-10 23:17:43	KYLIN_RHEOS_METRICS_QUERY_QA2	100%	2017-05-10 23:23:37 PST	5.87 mins	Action	
KYLIN_RHEOS_METRICS_QUERY_RPC_QA2 - 20170510110000_20170510170000 - STREAM - PDT 2017-05-10 23:11:43	KYLIN_RHEOS_METRICS_QUERY_RPC_QA2	100%	2017-05-10 23:20:39 PST	8.73 mins	Action	
KYLIN_RHEOS_METRICS_QUERY_CUBE_QA2 - 20170510110000_20170510170000 - STREAM - PDT 2017-05-10 23:13:43	KYLIN_RHEOS_METRICS_QUERY_CUBE_QA2	100%	2017-05-10 23:20:06 PST	6.12 mins	Action	
KYLIN_RHEOS_METRICS_QUERY_QA2 - 20170508110000_20170509110000 - MERGE - PDT 2017-05-10 23:12:58	KYLIN_RHEOS_METRICS_QUERY_QA2	100%	2017-05-10 23:17:18 PST	4.30 mins	Action	
KYLIN_RHEOS_METRICS_QUERY_CUBE_QA2 - 20170509110000_20170510110000 - MERGE - PDT 2017-05-10 23:09:10	KYLIN_RHEOS_METRICS_QUERY_CUBE_QA2	100%	2017-05-10 23:13:40 PST	4.48 mins	Action	
KYLIN_RHEOS_METRICS_QUERY_RPC_QA2 -	KYLIN_RHEOS_METR	100%	2017-05-10	4.17 min	Action	

2017-05-10 23:49:45 PST

#1 Step Name: Build Dimension Dictionaries For Steaming Job
Data Size: 2.23 KB
Duration: 1.04 mins

2017-05-10 23:50:47 PST

#2 Step Name: Save Cube Dictionaries
Duration: 0.01 mins

2017-05-10 23:50:48 PST

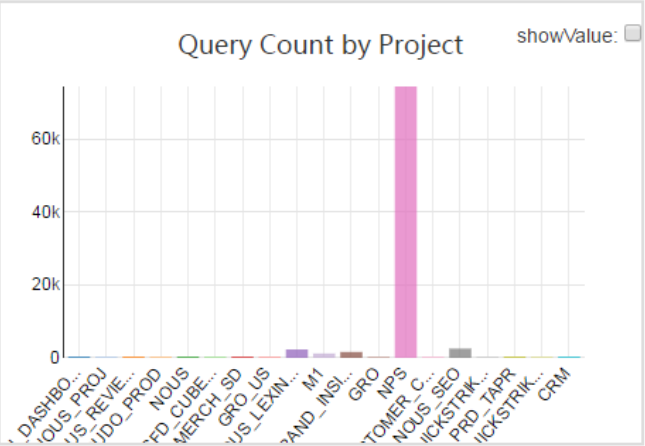
#3 Step Name: Build Base Cuboid Data For Streaming Job
Data Size: 4.22 KB
Duration: 1.08 mins

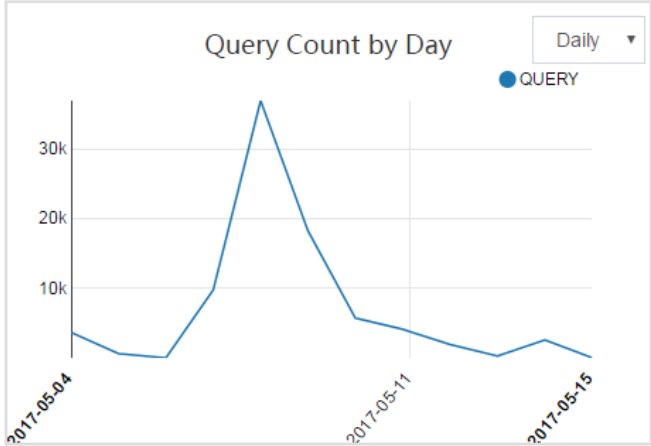
Use Case

ebay CHINA CENTER OF EXCELLENCE

TOTAL CUBE COUNT	QUERY COUNT	AVG QUERY LATENCY	JOB COUNT	AVG BUILD TIME PER MB
11	84,023	0.15 s	223	7.00 sec
More Details	More Details	More Details	More Details	More Details

AVG CUBE EXPANSION
4.33
More Details





Thanks!