

Real-time Big Data in Real Life

Apache Kylin in production



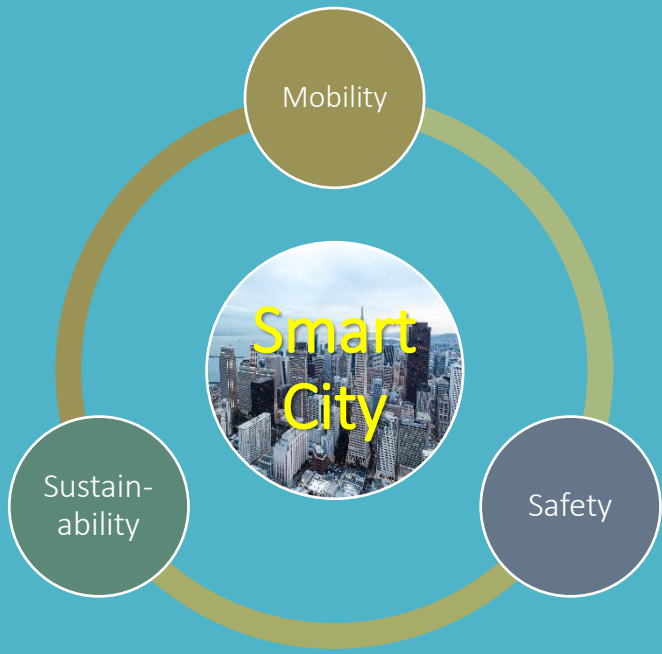
András István Nagy / EPAM

About me



Sequoia sempervirens

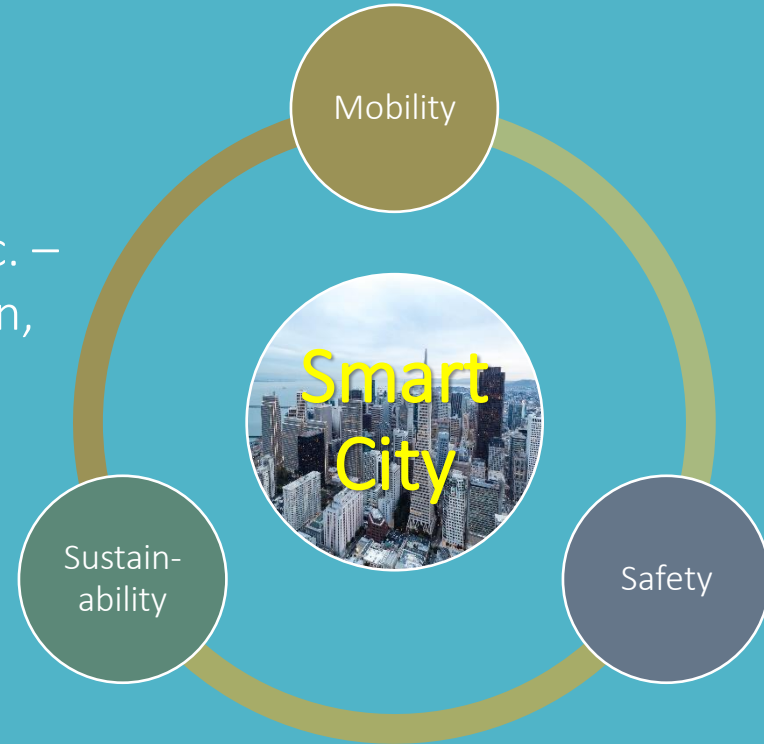
András

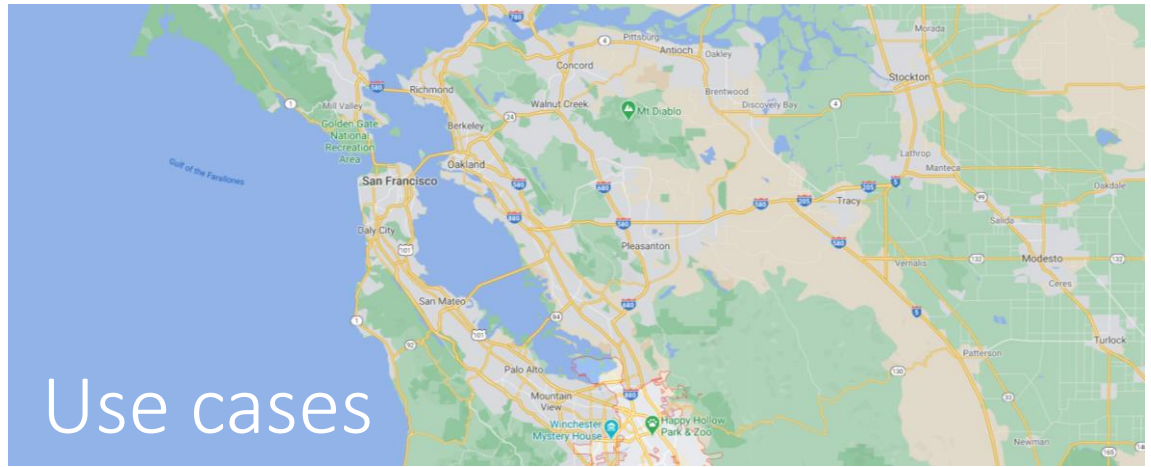


Smart City traffic use-cases

Traffic Throughput and Safety

- **Maximize traffic throughput**
 - Analyze traffic patterns at intersections
 - Vehicle count, speed, queue length, etc. – breakdown by vehicle class, intersection, time periods, direction, lane etc.
 - Volume-Capacity Ratio
- **Increase Safety**
 - Detect dangerous patterns:
 - Near misses
 - Jaywalking persons
 - Yellow-light dilemma zone





Batch analytics - for Traffic Engineers:

- Optimize traffic lights, lanes
- Optimize location of fire stations

Real-time – intervention & automation:

- Service car to check up on unexpected traffic events
- Avoid yellow-light dilemma
- Emergency vehicle routing
- The future is CAV



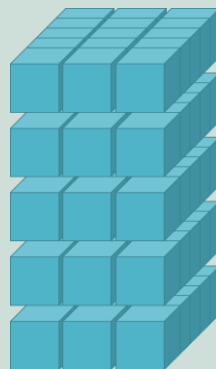
Traffic sign providers



IoT devices



Master data



Data Processing & Analytics

Insight



- Real-time route planning



- Real-time Dashboards for traffic engineers



- Real-time traffic light control



- Location optimization (emergency vehicles)

Analytics Data Flow


```
{  
  "intersection": "I-079"  
  "vehicle": {  
    "type": "car",  
    "velocity": "23",  
    "latitude": "41.88775700568241",  
    "longitude": "-87.62228626863555",  
  }  
  "signal_phase": "red"  
  "timestamp": "2020-09-19T18:05:03.122",  
}
```






Edge processing

~ 100k - 300k events / intersection / day

Interfaces
UIs and APIs

 Machine Learning

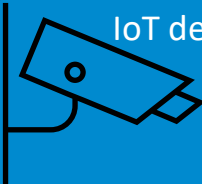
Kafka



Kylin
real-time
aggregations

Data
Lake

Event pre-
processing
Spark Streaming /
Microservices

 IoT devices

Other sources



Why do we need a specialized tool?

Previous approaches:

- On-the-fly queries with Presto

(Drill/Impala/etc.)

- not fast enough for long periods
- **NOT REAL TIME** performance

- Custom data aggregation jobs

- Very similar implementations for each use-case
- **TOO MUCH** custom code
- Efficient incremental aggregation is not trivial
- **CUSTOM CODE** makes long development iterations, difficult to achieve real-time feedback

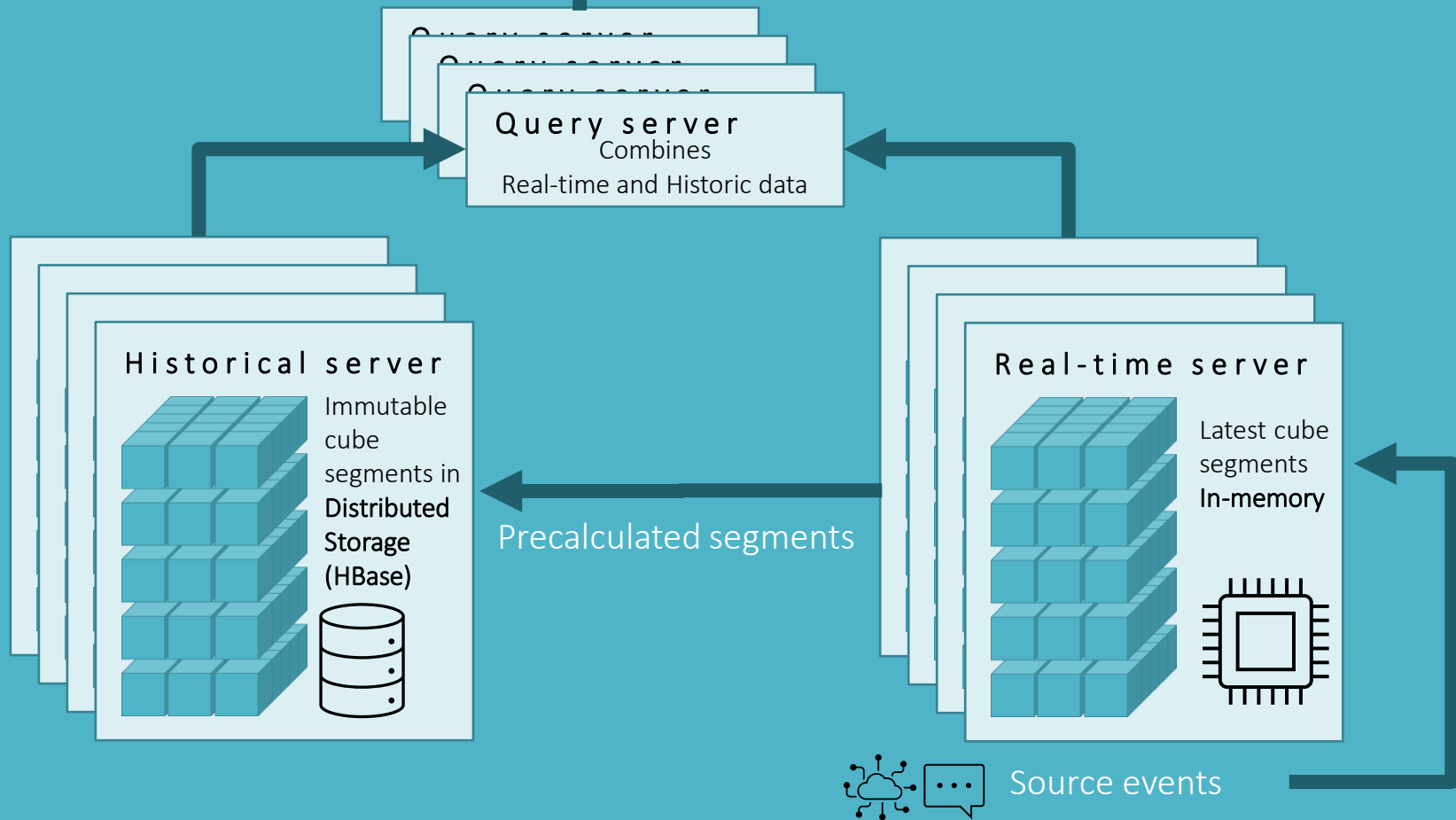


Why Kylin?

- Distributed system that can scale to streaming Big Data
- Maintainability
- Maturity
- Performance
- Manageability (operations)
- Interfacing with existing architecture
- High Availability
- Enterprise Security compliance
- Licence



Visualizations



All of this in
production...



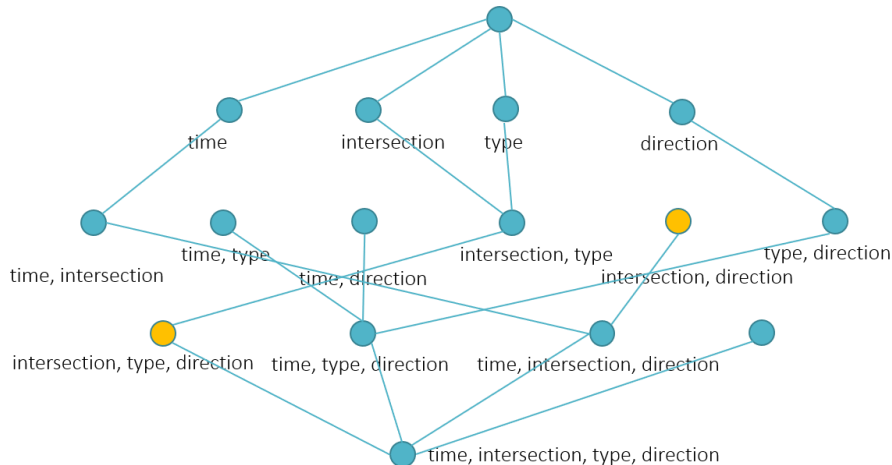
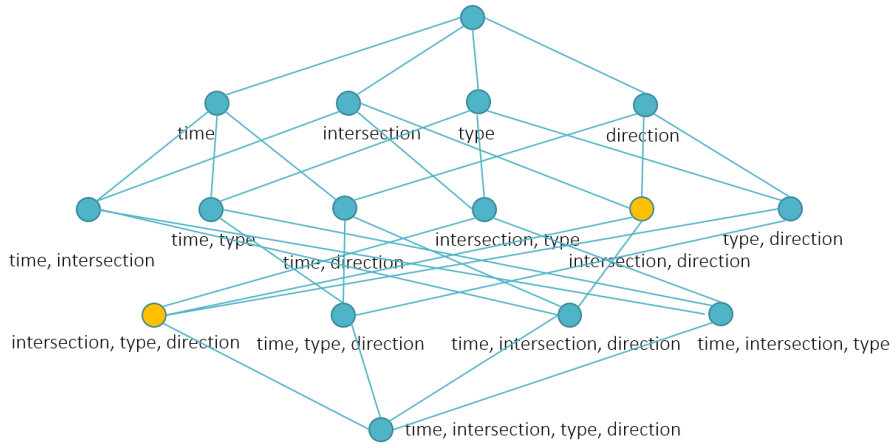
Streaming aggregations in real life

IoT Reality: Late events

- Slightly late events:
 - Kylin keeps in-memory segments open if events keep coming in for the time period
- Very late events:
 - Lambda mode – use a Hive table with the same schema as the Kafka events
 - Can rebuild previous day in batch → orchestration on rebuilds
- Tradeoff between slightly late and very late - Use-case specific solution design

Note: Late event support is a question of the end-to-end pipeline

Large number of dimensions



Cube optimizations

- Derive cuboid from parent with smallest size
- Mandatory dimensions
- Aggregation groups

Cube optimizer for interactive use

- Learn about usage patterns and suggest optimizations

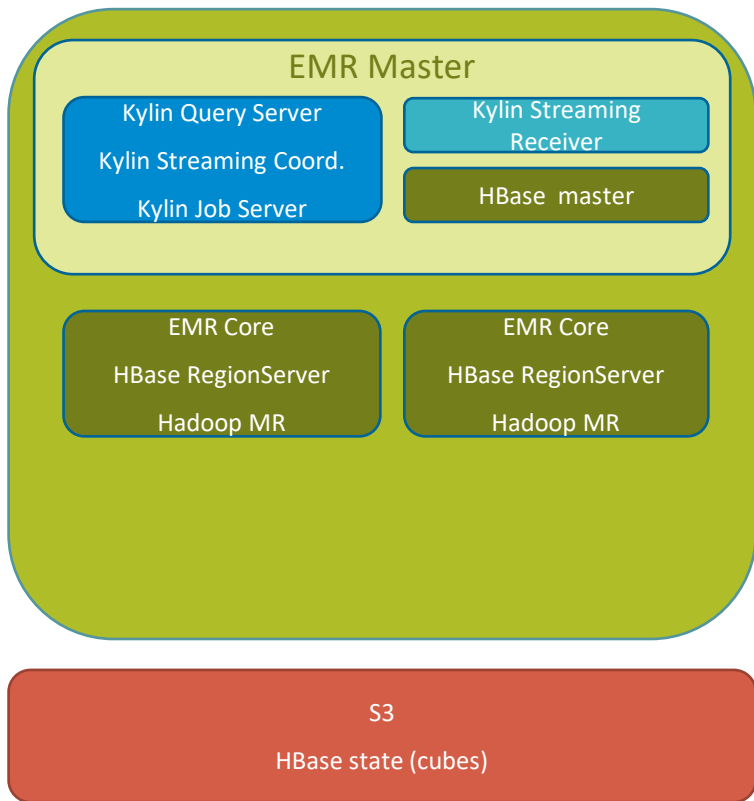
Hosting & Operations

- Real-time OLAP on Big Data as **managed cloud service?**
 - *Not aware of any...*
 - Apache Druid is part of Azure HDInsight
 - *Still not a managed service...*
 - Kyligence Cloud available as a cloud offering with support on AWS, Azure, GCP, AlibabaCloud
 - *Still not a managed service...*

Still need:

- Infrastructure provisioning automation
 - Terraform / CloudFormation / ...
- Cube deployment pipeline
 - DEV → TEST → PROD
- Monitoring and alerting

AWS Deployment – PoC setup



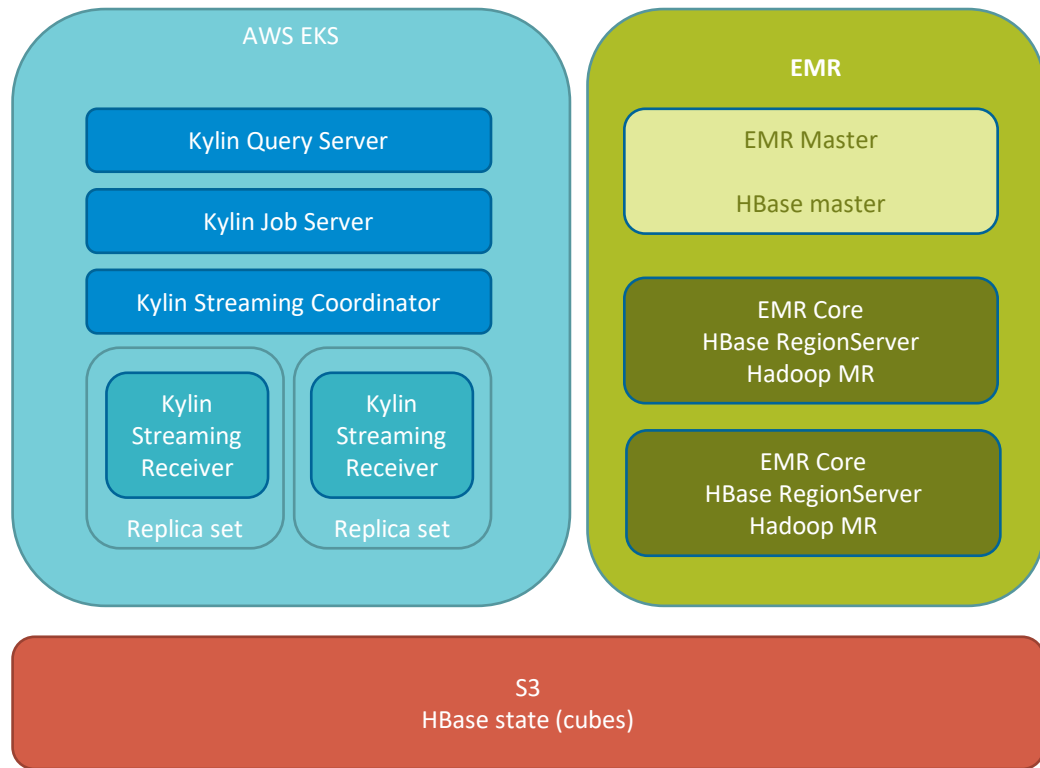
Design goal: **stateless EMR clusters**

- Persistent data in HBase → S3
- Persistent metadata in Hive Metastore → Aurora

Infrastructure Automation

- Terraform

AWS Deployment – Initial PROD setup



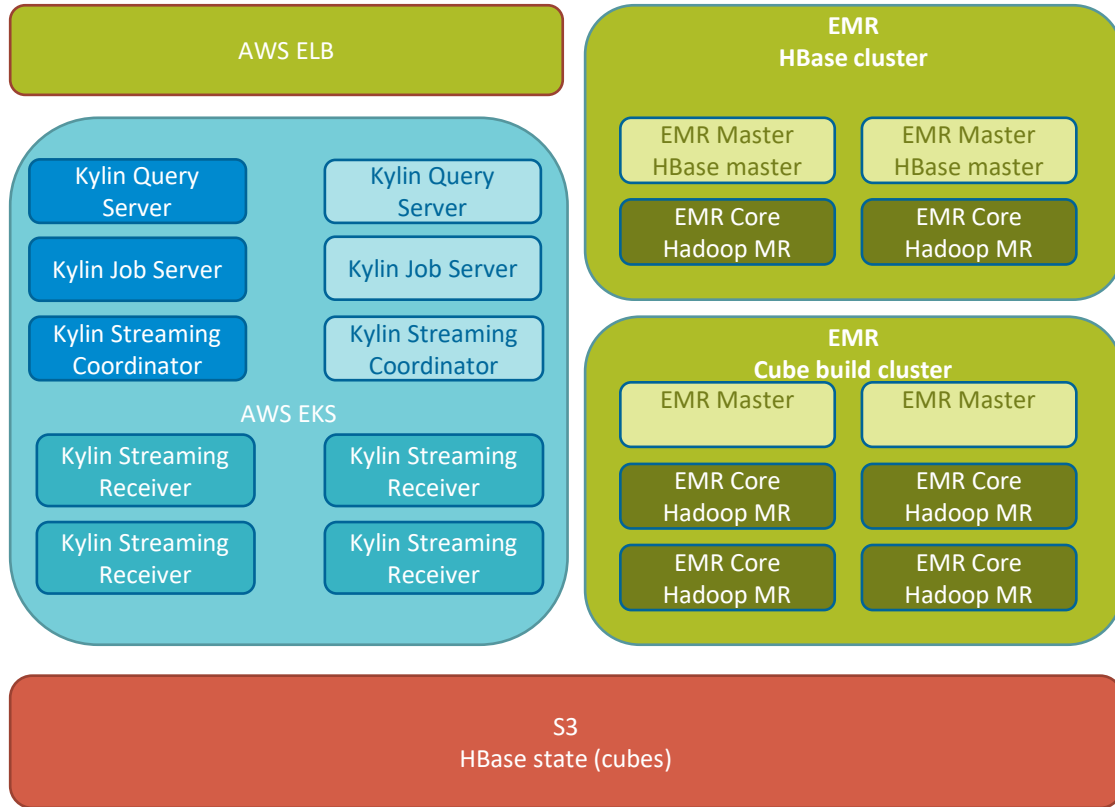
Design goal: **stateless EMR clusters**

- Persistent data in HBase → S3
- Persistent metadata in Hive Metastore → Aurora

Infrastructure Automation

- Terraform
- Separated Kylin components from EMR master

AWS Deployment – Scaled & HA PROD setup



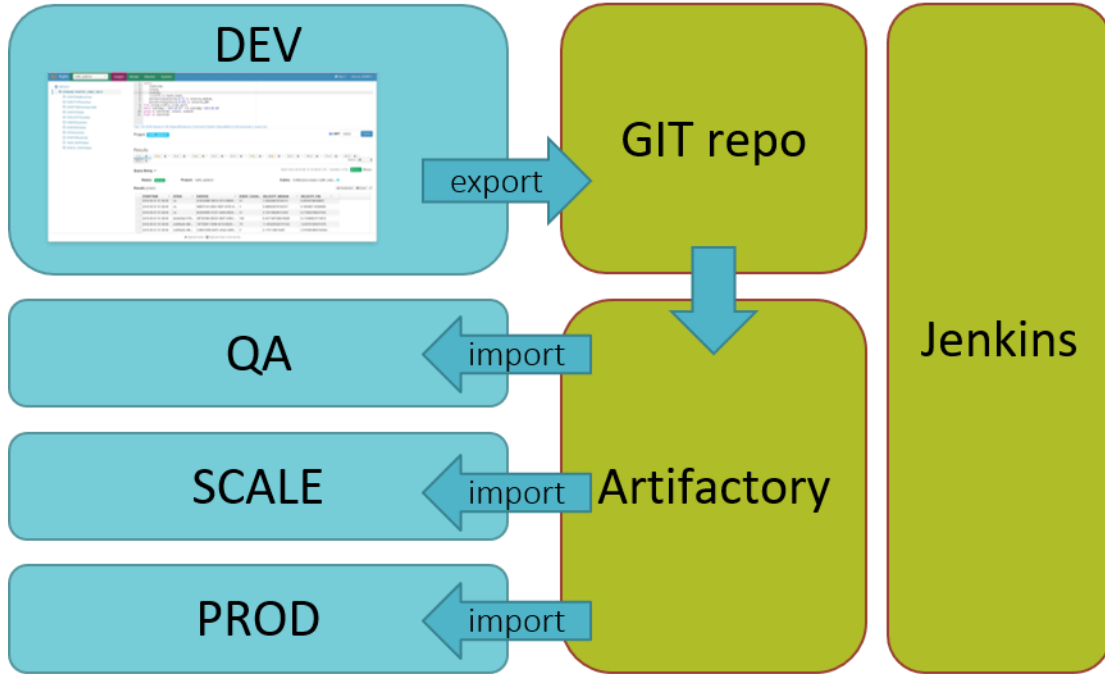
Design goal: **stateless EMR clusters**

- Persistent data in HBase → S3
- Persistent metadata in Hive Metastore → Aurora

Infrastructure Automation

- Terraform
- Separate EMR clusters for HBase and cube building
- All Kylin components replicated & HA

Cube deployment pipeline



Cubes are not defined by code but assembled on a UI

How does this fit into a SDLC?

- Builds / releases
- Dev / test / prod environments
- Roll out tested releases to PROD

Apache Kylin | OLAP engine for ...

Not Secure | kylin.apache.org

Home Docs Download Community Development Blog 中文页

Who are using Kylin?

Kylin Ecosystem

www.stratebi.com

Apache Kylin | Community

Not secure | kylin.apache.org/community/

Home Docs Download Community Development Blog 中文页

Apache Kylin™

Extreme OLAP Engine for Big Data

Powered By Apache Kylin™

For information about who are using Apache Kylin™, please refer to [Powered By](#) page.

Apache Kylin Mailing List

There are the mailing lists that have been established for this project. For each list there is a subscribe, unsubscribe, and an archive link. Note: if you do not receive the confirmation email after sending email to the mail list, the email message is shown in your trash mail.

User Mailing List [Subscribe](#) [Unsubscribe](#) [Post mail-archives.apache.org](#)

Developers Mailing List [Subscribe](#) [Unsubscribe](#) [Post mail-archives.apache.org](#)

Issues Mailing List [Subscribe](#) [Unsubscribe](#) [N/A](#) [mail-archives.apache.org](#)

Commith Mailing List [Subscribe](#) [Unsubscribe](#) [N/A](#) [mail-archives.apache.org](#)

Community Activity Report

GitHub

Social Media

The official Kylin Twitter account: [@kylinchina](#)

Events and Conferences

Events

Apache Kylin Media [@kylinchina](#)

Apache Kylin Media [@kylinchina](#)

Process a list

Conferences

Accelerate Big Data Analytics with Apache Kylin

Refactor your data warehouse with mobile analytics products

More Events and Conferences

Mailing List Archives

For convenience, there's a forum style mailing list archives which not part of official Apache archives:

Takeaways

Big Data tools moving towards **higher levels of abstraction** and **real-time analytics**

Kylin **replaces a lot of custom development**

Efficient calculations with **intelligent** optimizations

You will still need a **lot of automation**

Move all state to **PaaS storage** 😊