十年架构 成长之路

SACC 第十届中国系统架构师大会

SYSTEM ARCHITECT CONFERENCE CHINA 2018

2018年10月17-10月21日 北京海淀永泰福朋喜来登酒店

IT168.com ChinaUnix ITPUB

Apache Griffin

Data Quality Solution for both streaming and batch

郭跃鹏

eBay资深主任工程师
数据服务方案部门
guoyp@apache.org

Agenda

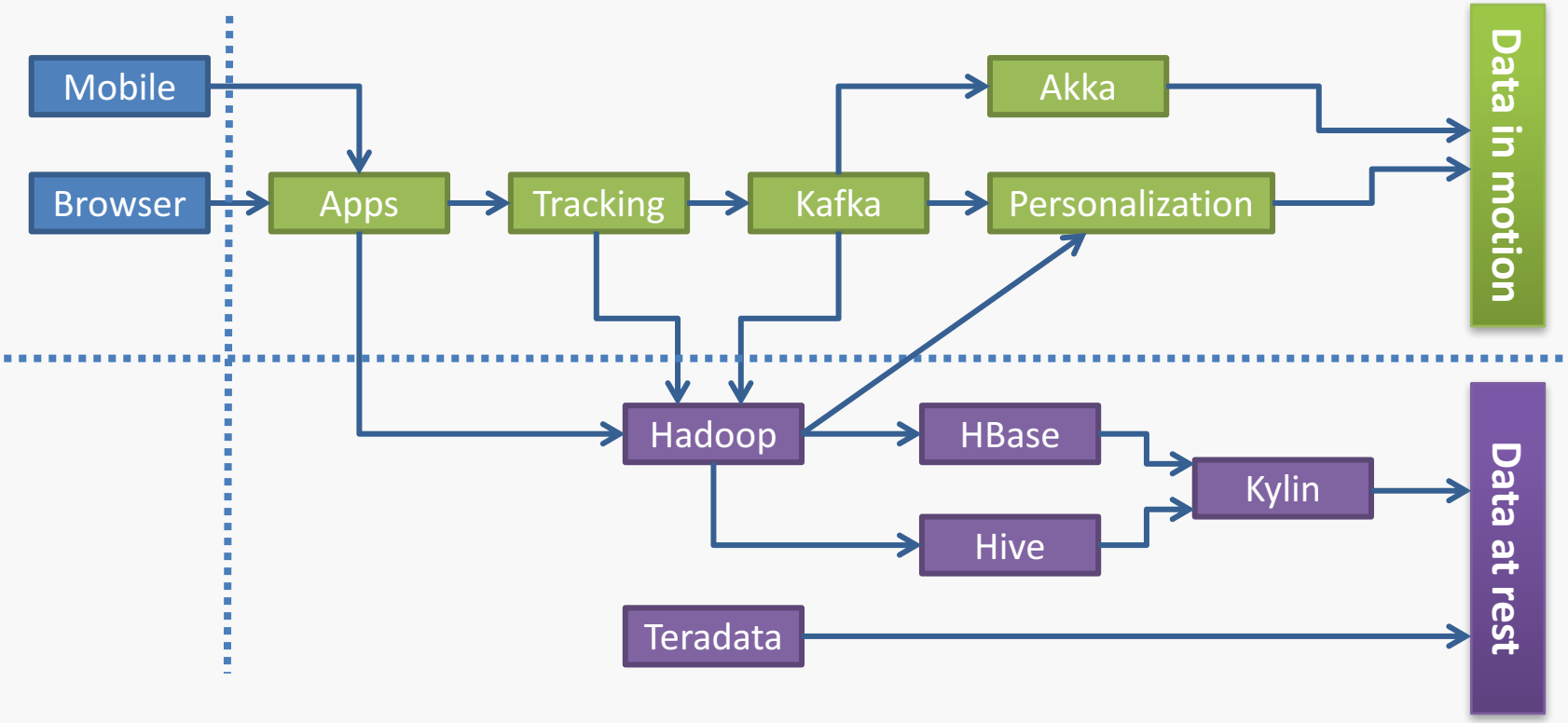
- Background
- What is Apache Griffin
- How
- New Features
- What is Next
- How to Contribute
- Q/A

Background

One day ,personalization team found a large decrease in data quality metrics

Date	Schema Name	Accuracy Rate	Target Rate
2016-02-13	viewitem	93.42%	99.99%
2016-02-13	search	77.13%	99.99%
2016-02-13	bid_new	98.66%	99.99%
2016-02-13	transaction_new	100.0%	99.99%
2016-02-13	item_watch	96.08%	99.99%

Background


Mobile
Browser
Apps
Tracking
Kafka
Akka
Personalization
Data in motion
Hadoop
HBase
Kylin
Hive
Teradata
Data at rest

Background

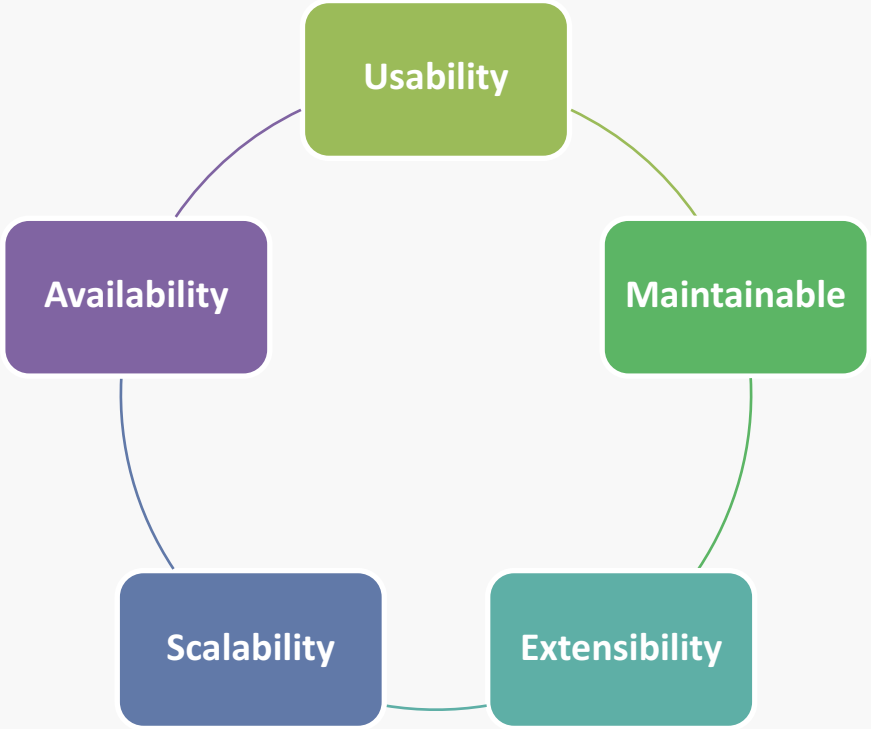
- No **unified** view of data quality across multiple systems and teams
- No **shared** platform to manage data quality
- No **near real-time** report for system health status

What is Apache Griffin

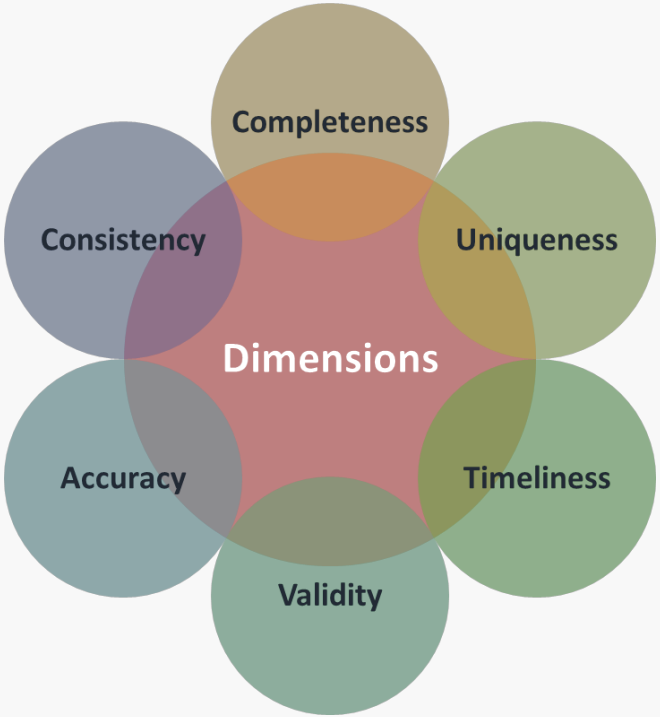
- A unified Data Quality Platform
- A unified process to detect various DQ issues
- An open source solution

https://github.com/apache/incubator-griffin

What is Apache Griffin

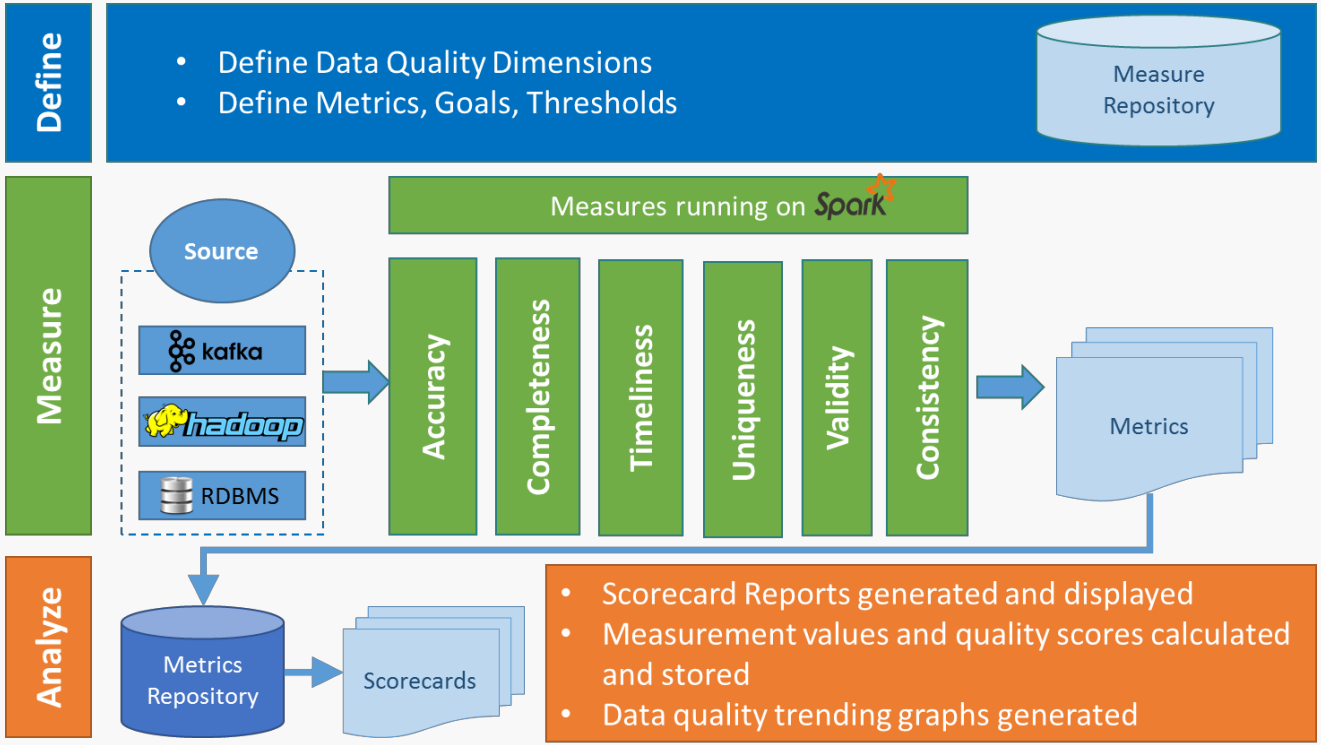


What is Data Quality

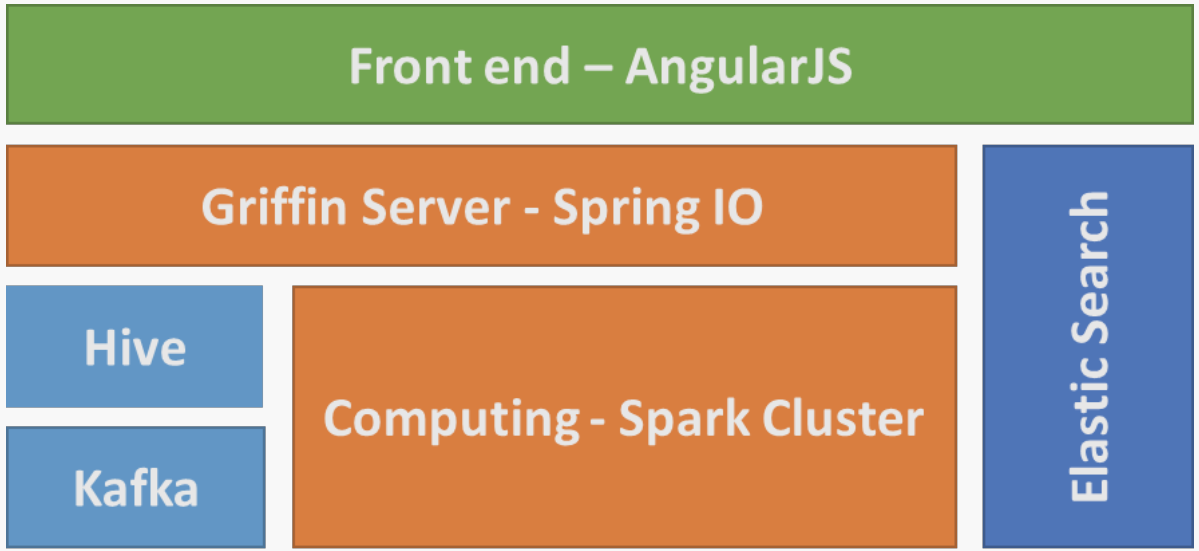


Data Quality Process Lifecycle

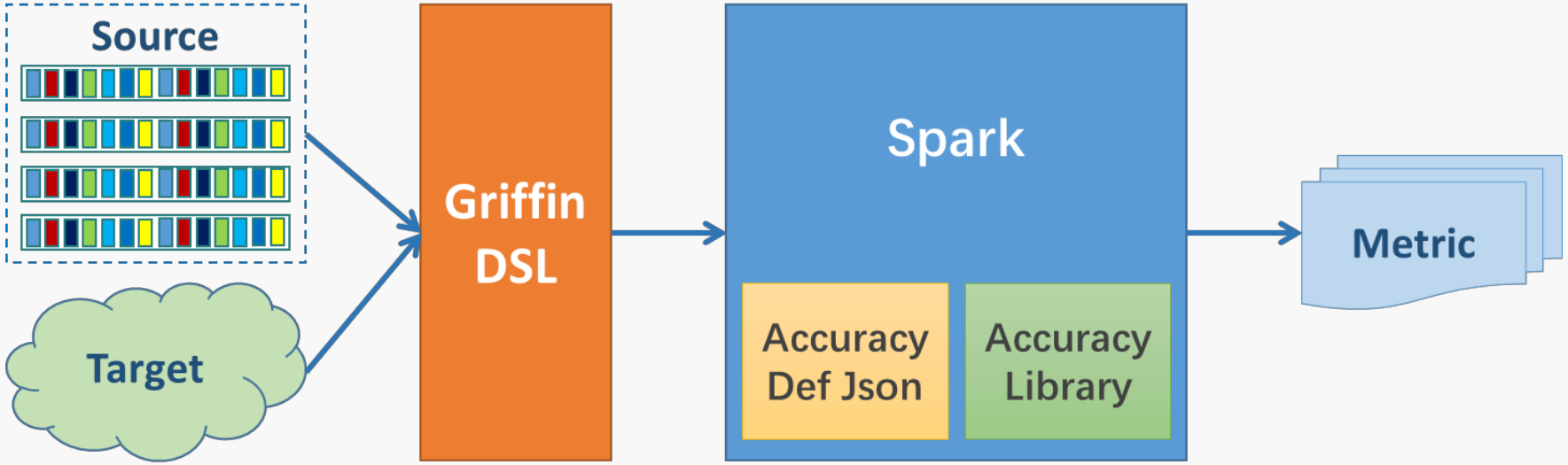
Apache Griffin Architecture



Apache Griffin Tech Stack



How - Batch Mode



$$Accuracy\ Rate(\%) = \frac{Count(source.field1 == target.field1\ \&\&\ \dots)}{Count(source)}\ X\ 100\%$$

How - Batch Mode

```
{
    “process.type”:”batch”
    “data.sources”:[
        {“name”:”source”, “connector”:”TableA”}
        {“name”:”target”, “connector”:”TableB”}
    ]
    “rules”:[
        {“rule”:”source.userid = target.userid AND …”}
    ]
    ”sinks”:[
        “LOGS”,”ELASTICSEARCH”
    ]
}
```

How - Streaming Mode

```
{
    “process.type”:”streaming”
    “data.sources”:[
        {“name”:”source”, “connector”:”TopicA”, type: “kafka”,
         ”interval”:”10s”, “time.range”:[-2m,0]}
        {“name”:”target”, “connector”:”TopicB”, type: “kafka”,
        ”interval”:”10s”, “time.range”:[-2m,0]}
        }
    ]
    “rules”:[
        {“rule”:”source.userid = target.userid AND …”}
    ]
    ”sinks”:[
        “LOGS”,”ELASTICSEARCH”
    ]
}
```

What's new

- Streaming supported
- Based on Spark 2.2.0
- New DSL

What’s next

- More connectors
- Topologic based data quality

How to contribute

We are open source and PR/review are welcomed

GitHub : https://github.com/apache/incubator-griffin

Website : https://griffin.incubator.apache.org

Contact: mailto://subscribe-dev@griffin.incubator.apache.org

Apache Griffin JIRA: https://issues.apache.org/jira/browse/GRIFFIN

Apache Griffin Wiki : https://cwiki.apache.org/confluence/display/GRIFFIN/Griffin

Demo

Batch accuracy: https://griffin.incubator.apache.org/docs/quickstart.html

Batch profiling: https://griffin.incubator.apache.org/docs/profiling.html

Streaming accuracy : https://griffin.incubator.apache.org/docs/usecases.html

Q/A

THANKS