

DTCC

2017第八届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2017



# Apache Griffin

Data Quality Solution for both streaming and batch

郭跃鹏

eBay资深主任工程师

数据服务部门

guoyp@apache.org

# Agenda

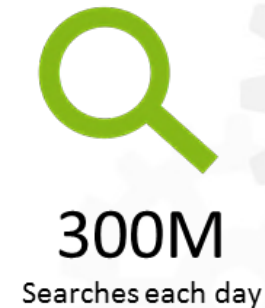
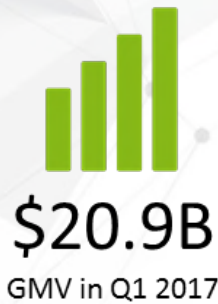
- About us
- Apache Griffin
- Demo
- What is coming
- How to contribute
- Q/A



# eBay Marketplace at a Glance



One of the world's largest and most vibrant marketplaces



# Velocity Stats



US

3 car parts or accessories are sold every	1 sec
A smartphone is sold every	4 sec
A dress is sold every	6 sec



UK

A make-up product is sold every	3 sec
A necklace is sold every	10 sec
A Lego product is sold every	19 sec



GERMANY

A truck or car is sold every	5 min
A pair of women's jeans is sold every	4 sec
A video game is sold every	11 sec



AUSTRALIA

A pair of men's sunglasses is sold every	1 min
A home décor item is sold every	12 sec
A car or truck part is sold every	4 sec



# Mobile Velocity Stats



US

A woman's handbag is sold every **10 sec**



A car or truck is sold every **5 min**

An action figure is sold every **10 sec**



UK

A cookware item is sold every **6 sec**



A tablet is sold every **1 min**

A car is sold every **2 min**



GERMANY

A pair of women's shoes is sold every **20 sec**



A watch is sold every **48 sec**

A tire or car part is sold every **35 sec**



AUSTRALIA

A piece of jewelry is sold every **12 sec**

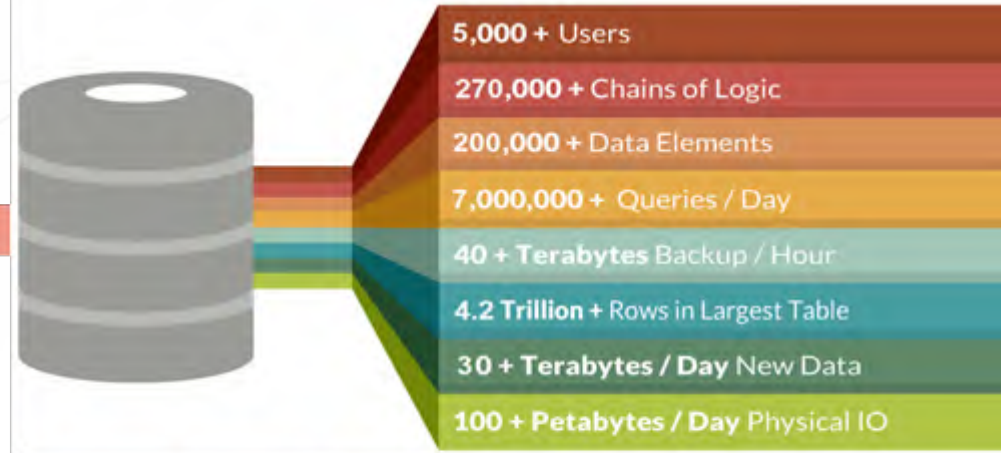


A baby clothing item is sold every **46 sec**

A motorcycle part is sold every **51 sec**



# Big Data @ ebay

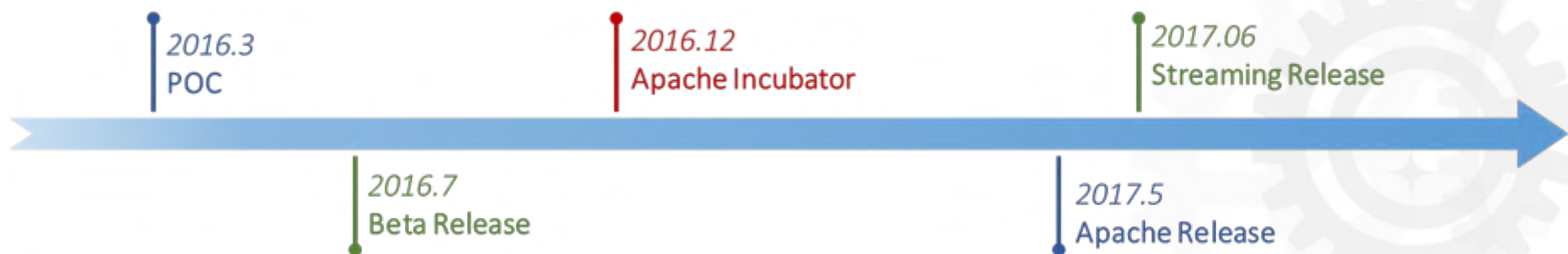


We utilize one of the largest data platforms in the world

# Apache Griffin



- Platform approach
- Common data quality dimensions
- Extensible, pluggable, scalable
- Trusted datasets



# A story - problem



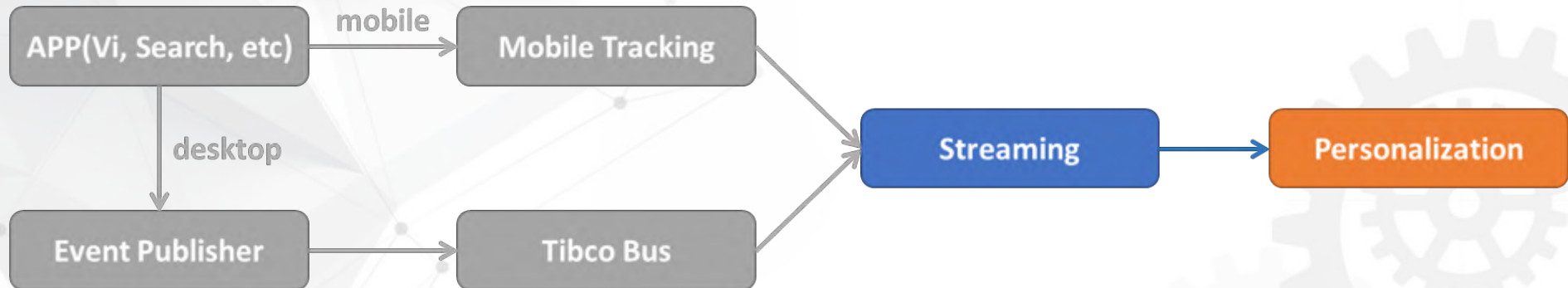
One day, personalization team found a large decrease in data quality metrics

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%





# A story - analysis



For large decrease in metrics, **candidates** include:

- The report are broken after streaming transfer due to minor changes in fields
- We are missing data from pipeline
- Our data queue is not working

P

We **didn't change anything**.

Hey Streaming team: Could you check from your side for this issue?

P

S

Let's check our metrics.

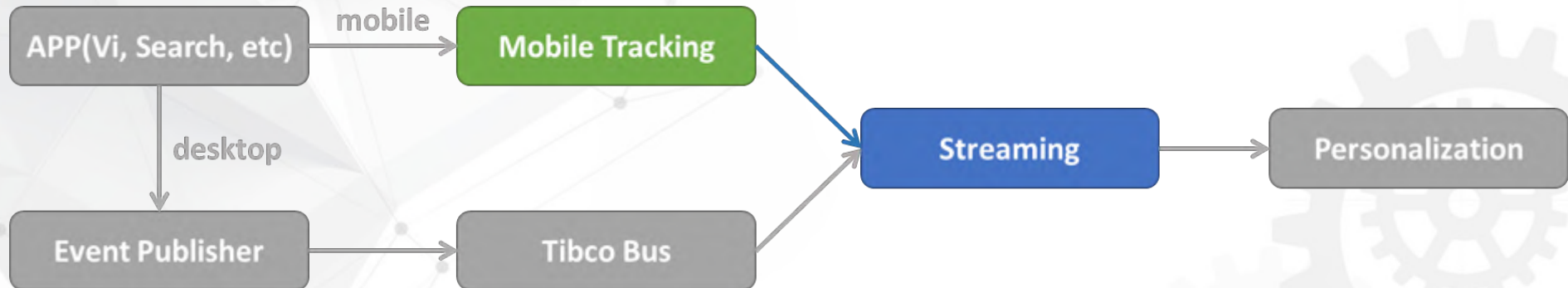
S

Oh... ok, we need to **add a new metrics** for this?

S

uh,... it seems data are already lost, let me check with our **upper streaming**

# A story – analysis continued



Hi Mobile,  
Can we temporarily switch/**restore** to old version?

S

M What is your logic for data quality from your side, show us your sql...

Select \* from ... a left outer join b on ...  
where... and ...

S

M uh, ...

...

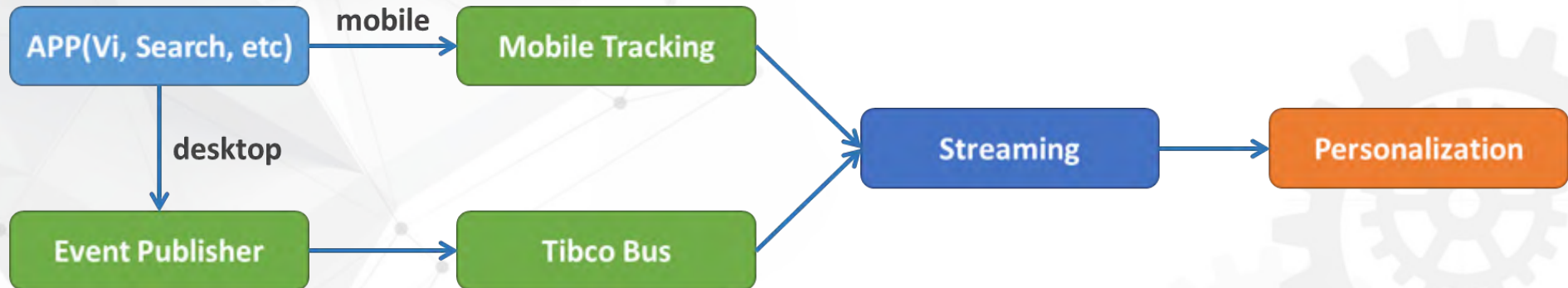
S

M **mobiles app will never send rq in version 4.1.5**

**Right! That's root cause.**

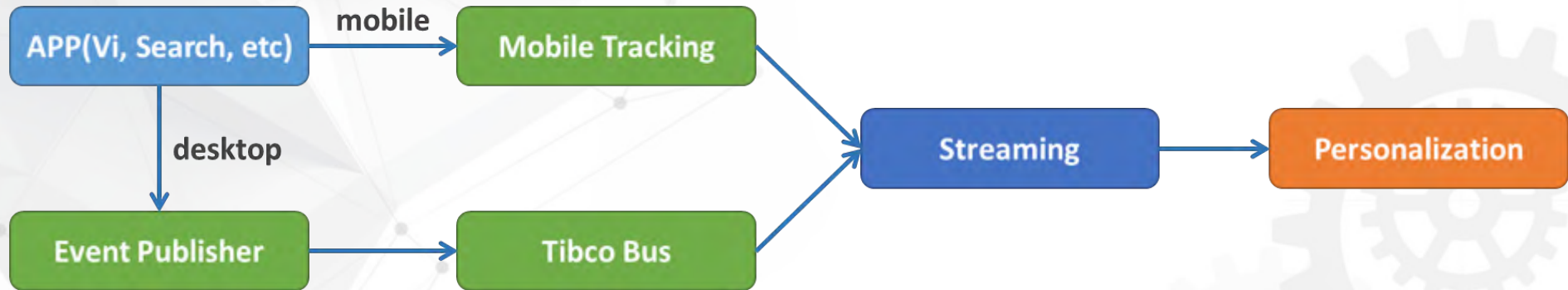
S

# A story – analysis continued



- **Isolated** system looks good from their own perspective.
- **Communication** is always hard when crossing teams.
- We took **1 week** to find the root cause.

# A story - conclusion



- No **unified** view of data quality across multiple systems and teams
- No **platform approach** to manage data quality
- No systematic way to measure **near real-time** data quality



# Apache Griffin



- **Data Quality Platform built on Hadoop and Spark**
  - Batch data
  - Real-time data
- **A unified process to detect DQ issues**
  - Inaccurate
  - Incomplete
  - Invalid
  - .....
- **An open source solution**

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

# Griffin Goal



A solution with all the below capabilities

Capability	Commercial DQ software	Open source DQ software	Apache Griffin
Support eBay's scale	x	x	√
Data Quality measurement	√	x	√
Support real-time data	x	x	√
Support unstructured data	x	x	√
Service based API	√	x	√
Data Profiling	√	√	√
Pluggable measurement types	x	x	√

# What is Data Quality?



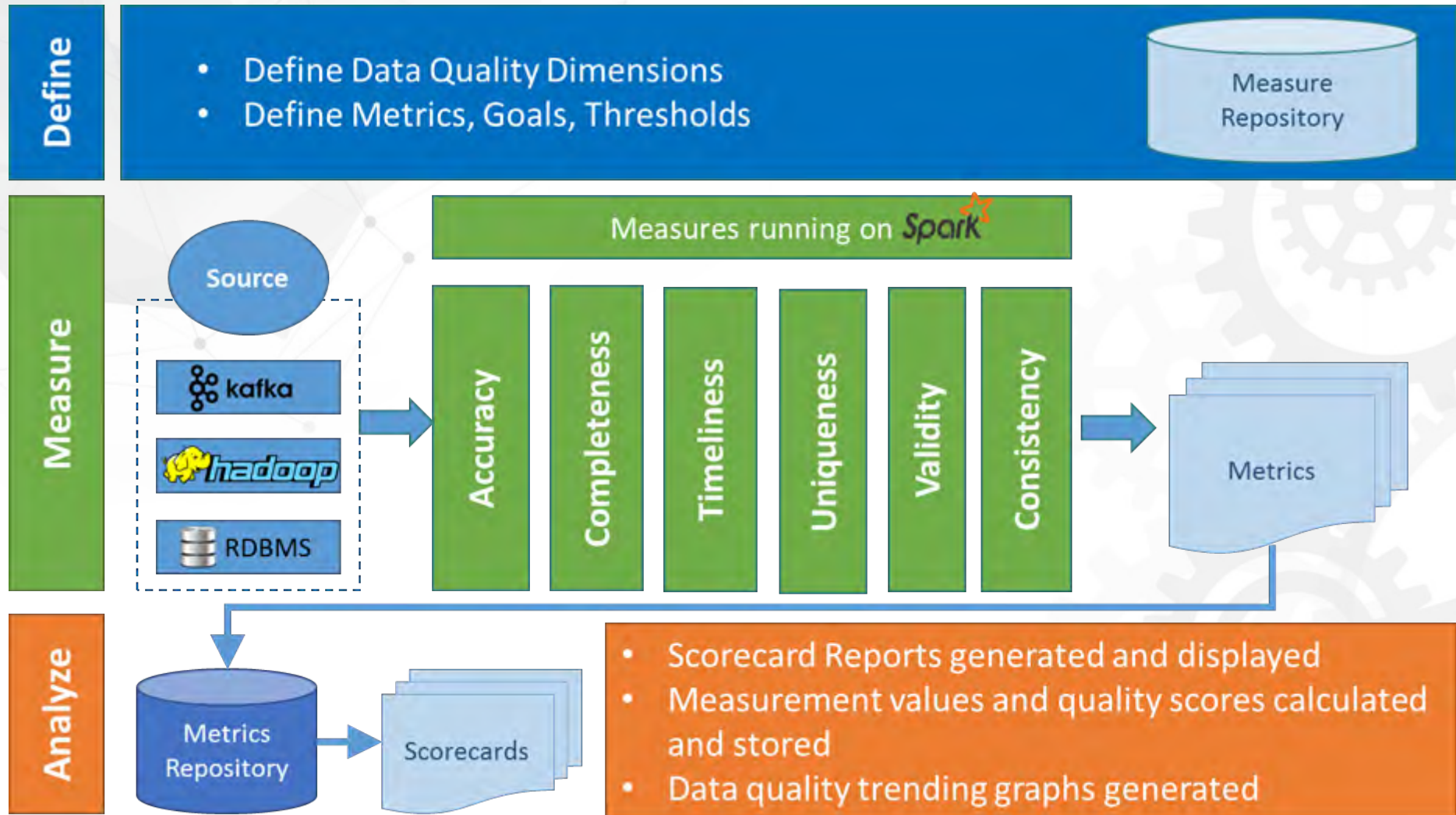
# Virtuous Cycle of Data Quality

- Define the scope, dimensions, goals, thresholds, etc.
- Measure data quality values
- Analyze data quality results
- Improve data quality





# Apache Griffin Architecture



# Apache Griffin – Tech Stack



Front end – AngularJS

Griffin Server - Spring IO

Hive

Kafka

Computing - Spark Cluster

Elastic Search

# Apache Griffin – Measure insights GriFFin

- Uniform Data Quality DSL
- DSL support both streaming and batch
- Configurable data source connectors

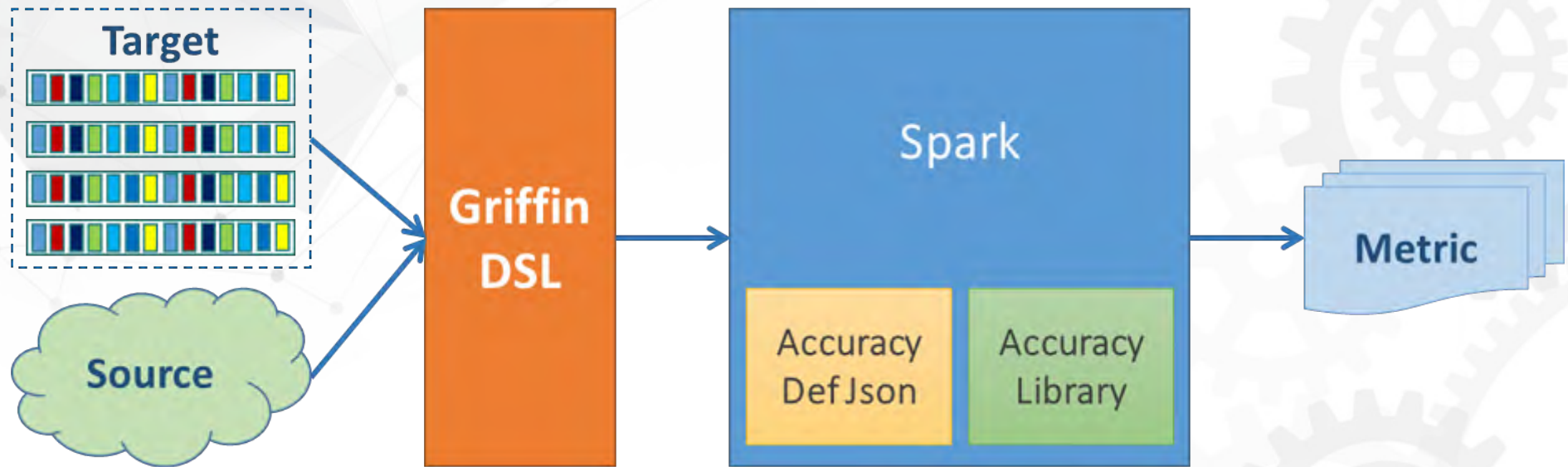
Accuracy DSL example:

Where: “\$source.uid = \$target.uid  
and \$source.itemid = \$target.itemid  
and \$source.tmp > \$target.tmp”

# Apache Griffin – Accuracy Measure



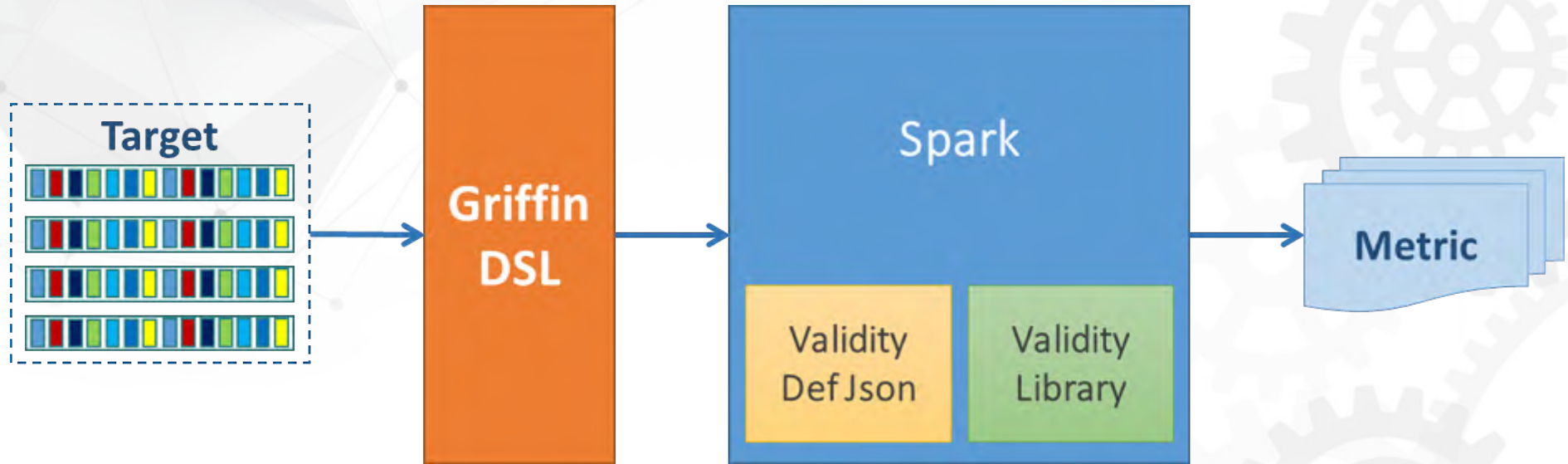
~300M customer view item events per day



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



# Apache Griffin – Validity Measure



# Apache Griffin – Time Series Metrics GriFFin

Elasticsearch :

- Offer aggregations
- Visualization(kibana, Grafana)
- Restful to integrate with

# Apache Griffin



Life is easier after Griffin ...



# Apache Griffin – Tech Challenges



- Unified model for both streaming and batch
- Stable
- Easily adaption
- Scalable, extensible algorithms



# Demo



2017年第八届中国数据库技术大会  
DATABASE TECHNOLOGY CONFERENCE CHINA 2017



# Use Cases

Griffin has been deployed in production at eBay and provided the centralized data quality service for several eBay systems.

**Data**

~1.2PB



**Daily Records**

800+M



**Metrics**

100+



# What is coming



- DSL to support more dimensions
  - ✓ Completeness
  - ✓ Consistency
  - ✓ Anomaly detection
- Provide more data source connectors
  - ✓ Raw Hadoop data
  - ✓ Hybrid data source connectors

# How to Contribute



- Community over code
- Meritocracy

# How to Contribute



We are open source and PR 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>





Q / A



DTCC

2017年第八届中国数据库技术大会  
DATABASE TECHNOLOGY CONFERENCE CHINA 2017

SequeMedia  
数据传媒

IT168

mpub

ChinaUnix



# THANKS

SequeMedia  
盛拓传媒

IT168.com

ITPUB

ChinaUnix