GriffinProposal

Griffin Proposal

Abstract

Griffin is a Data Quality Service platform built on Apache Hadoop and Apache Spark. It provides a framework process for defining data quality model, executing data quality measurement, automating data profiling and validation, as well as a unified data quality visualization across multiple data systems. It tries to address the data quality challenges in big data and streaming context.

Proposal

Griffin is a open source Data Quality solution for distributed data systems at any scale in both streaming or batch data context. When people use open source products (e.g. Apache Hadoop, Apache Spark, Apache Kafka, Apache Storm), they always need a data quality service to build his/her confidence on data quality processed by those platforms. Griffin creates a unified process to define and construct data quality measurement pipeline across multiple data systems to provide:

- Automatic quality validation of the data
- Data profiling and anomaly detection
- Data quality lineage from upstream to downstream data systems.
- Data quality health monitoring visualization
- Shared infrastructure resource management

Overview of Griffin

Griffin has been deployed in production at eBay serving major data systems, it takes a platform approach to provide generic features to solve common data quality validation pain points. Firstly, user can register the data asset which user wants to do data quality check. The data asset can be batch data in RDBMS (e.g.Teradata), Apache Hadoop system or near real-time streaming data from Apache Kafka, Apache Storm and other real time data platforms. Secondly, user can create data quality model to define the data quality rule and metadata. Thirdly, the model or rule will be executed automatically (by the model engine) to get the sample data quality validation results in a few seconds for streaming data. Finally, user can analyze the data quality results through built-in visualization tool to take actions.

Griffin includes:

Data Quality Model Engine

Griffin is model driven solution, user can choose various data quality dimension to execute his/her data quality validation based on selected target data-set or source data-set (as the golden reference data). It has a corresponding library supporting it in back-end for the following measurement:

- Accuracy - Does data reflect the real-world objects or a verifiable source
- Completeness - Is all necessary data present
- Validity - Are all data values within the data domains specified by the business
- Timeliness - Is the data available at the time needed
- Anomaly detection - Pre-built algorithm functions for the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset
- Data Profiling - Apply statistical analysis and assessment of data values within a dataset for consistency, uniqueness and logic.

Data Collection Layer

We support two kinds of data sources, batch data and real time data.

For batch mode, we can collect data source from Apache Hadoop based platform by various data connectors.

For real time mode, we can connect with messaging system like Kafka to near real time analysis.

Data Process and Storage Layer

For batch analysis, our data quality model will compute data quality metrics in our spark cluster based on data source in Apache Hadoop.

For near real time analysis, we consume data from messaging system, then our data quality model will compute our real time data quality metrics in our spark cluster. for data storage, we use time series database in our back end to fulfill front end request.

Griffin Service

We have RESTful web services to accomplish all the functionalities of Griffin, such as register data asset, create data quality model, publish metrics, retrieve metrics, add subscription, etc. So, the developers can develop their own user interface based on these web services.

Background

At eBay, when people play with big data in Apache Hadoop (or other streaming data), data quality often becomes one big challenge. Different teams have built customized data quality tools to detect and analyze data quality issues within their own domain. We are thinking to take a platform approach to provide shared Infrastructure and generic features to solve common data quality pain points. This would enable us to build trusted data assets.

Currently it’s very difficult and costly to do data quality validation when we have big data flow across multi-platforms at eBay (e.g. Oracle, Apache Hadoop, Couchbase, Apache Cassandra, Apache Kafka, MongoDB). Take eBay real time personalization platform as an example. Every day we have to validate data quality status for ~600M records (imagine we have 150M active users for our website). Data quality often becomes one big challenge both in its streaming and batch pipelines.

So we conclude 3 data quality problems at eBay:

1. Lack of end2end unified view of data quality measurement from multiple data sources to target applications, it usually takes a long time to identify and fix poor data quality. 2. How to get data quality measured in streaming mode, we need to have a process and tool to visualize data quality insights through registering dataset which you want to check data quality, creating data quality measurement model, executing the data quality validation job and getting metrics insights for action taking. 3. No Shared platform and API Service, have to apply and manage own hardware and software infrastructure.

Rationale

The challenge we face at eBay is that our data volume is becoming bigger and bigger, system processes become more complex, while we do not have a unified data quality solution to ensure the trusted data sets which provide confidences on data quality to our data consumers. The key challenges on data quality includes:

1. Existing commercial data quality solution cannot address data quality lineage among systems, cannot scale out to support fast growing data at eBay 2. Existing eBay's domain specific tools take a long time to identify and fix poor data quality when data flowed through multiple systems 3. Business logic becomes complex, requires data quality system much flexible. 4. Some data quality issues do have business impact on user experiences, revenue, efficiency & compliance. 5. Communication overhead of data quality metrics, typically in a big organization, which involve different teams.

The idea of Griffin is to provide Data Quality validation as a Service, to allow data engineers and data consumers to have:

- Near real-time understanding of the data quality health of your data pipelines with end-to-end monitoring, all in one place.
- Profiling, detecting and correlating issues and providing recommendations that drive rapid and focused troubleshooting
- A centralized data quality model management system including rule, metadata, scheduler etc.
- Native code generation to run everywhere, including Hadoop, Kafka, Spark, etc.
- One set of tools to build data quality pipelines across all eBay data platforms.

Current Status

Meritocracy

Griffin has been deployed in production at eBay and provided the centralized data quality service for several eBay systems (for example, real time personalization platform, eBay real time ID linking platform, Hadoop datasets, Site speed analytics platform). Our aim is to build a diverse developer and user community following the Apache meritocracy model. We will encourage contributions and participation of all types of work, and ensure that contributors are appropriately recognized.

Community

Currently the project is being developed at eBay. It's only for eBay internal community. Griffin seeks to develop the developer and user communities during incubation. We believe it will grow substantially by becoming an Apache project.

Core Developers

Griffin is currently being designed and developed by engineers from eBay Inc. – William Guo, Alex Lv, Shawn Sha, Vincent Zhao, John Liu. All of these core developers have deep expertise in Apache Hadoop and the Hadoop Ecosystem in general.

Alignment

The ASF is a natural host for Griffin given that it is already the home of Hadoop, Beam, HBase, Hive, Storm, Kafka, Spark and other emerging big data products. Those are requiring data quality solution by nature to ensure the data quality which they processed. When people use open source data technology, the big question to them is that how we can ensure the data quality in it. Griffin leverages lot of Apache open-source products. Griffin was designed to enable real time insights into data quality validation by shared Infrastructure and generic features to solve common data quality pain points.

Known Risks

Orphaned Products

The core developers of Griffin team work full time on this project. There is no risk of Griffin getting orphaned since at least one large company (eBay) is extensively using it in their production Hadoop and Spark clusters for multiple data systems. For example, currently there are 4 data systems at eBay (real time personalization platform, eBay real time ID linking platform, Hadoop, Site speed analytics platform) are leveraging Griffin, with more than ~600M records for data quality status validation every day, 35 data sets being monitored, 50+ data quality models have been created.

As Griffin is designed to connect many types of data sources, we are very confident that they will use Griffin as a service for ensuring the data quality in open source data ecosystems. We plan to extend and diversify this community further through Apache.

Inexperience with Open Source

Griffin's core engineers are all active users and followers of open source projects. They are already committers and contributors to the Griffin Github project. All have been involved with the source code that has been released under an open source license, and several of them also have experience developing code in an open source environment. Though the core set of Developers do not have Apache Open Source experience, there are plans to onboard individuals with Apache open source experience on to the project.

Homogenous Developers

The core developers are from eBay. Apache Incubation process encourages an open and diverse meritocratic community. Griffin intends to make every possible effort to build a diverse, vibrant and involved community. We are committed to recruiting additional committers from other companies based on their contribution to the project.

Reliance on Salaried Developers

eBay invested in Griffin as a company-wide data quality service platform and some of its key engineers are working full time on the project. they are all paid by eBay. We look forward to other Apache developers and researchers to contribute to the project.

Relationships with Other Apache Products

Griffin has a strong relationship and dependency with Apache Hadoop, Apache HBase, Apache Spark, Apache Kafka and Apache Storm, Apache Hive. In addition, since there is a growing need for data quality solution for open source platform (e.g. Hadoop, Kafka, Spark etc), being part of Apache's Incubation community, could help with a closer collaboration among these four projects and as well as others.

Documentation

Information about Griffin can be found at https://github.com/eBay/griffin

Initial Source

Griffin has been under development since early 2016 by a team of engineers at eBay Inc. It is currently hosted on Github.com under an Apache license 2.0 at https://github.com/eBay/griffin . Once in incubation we will be moving the code base to apache git library.

External Dependencies

Griffin has the following external dependencies.

- Basic*
- JDK 1.7+
- Scala
- Apache Maven
- JUnit
- Log4j
- Slf4j
- Apache Commons

Hadoop

- Apache Hadoop
- Apache HBase
- Apache Hive

DB

- InfluxData

Apache Spark

- Spark Core Library

REST Service

- Jersey
- Spring MVC

Web frontend

- AngularJS
- jQuery
- Bootstrap
- RequireJS
- eCharts
- Font Awesome

Cryptography

Currently there's no cryptography in Griffin.

Required Resources

Mailing List

We currently use eBay mail box to communicate, but we'd like to move that to ASF maintained mailing lists.

Current mailing list: ebay-griffin-devs@googlegroups.com

Proposed ASF maintained lists:

private@griffin.incubator.apache.org

dev@griffin.incubator.apache.org

commits@griffin.incubator.apache.org

Subversion Directory

Git is the preferred source control system.

Issue Tracking

JIRA

Other Resources

The existing code already has unit tests so we will make use of existing Apache continuous testing infrastructure. The resulting load should not be very large.

Initial Committers

- William Go
- Alex Lv
- Vincent Zhao
- Shawn Sha
- John Liu
- Liang Shao

Affiliations

The initial committers are employees of eBay Inc.

Sponsors

Champion

Henry Saputra(hsaputra@apache.org)

Nominated Mentors

Kasper Sørensen(kaspersor@apache.org)

Uma Maheswara Rao Gangumalla(umamahesh@apache.org)

Luciano Resende(luckbr1975@gmail.com)

Sponsoring Entity

We are requesting the Incubator to sponsor this project.