BUILDING A SELF-SERVICE DATA PIPELINE WITH APACHE SPARK

ANDREW GRANGAARD

Created: 2018-03-10 Sat 12:41

We're Helping People Find Great Jobs, and Helping Employers Build Great Companies

- Scrappy, Data Driven, Experimental
  - Crawl, Walk, Run!
- Explosive Growth

BUSINESS GOAL:

Support Data Driven decision Making

- Collect **All** the data
- Ad-hoc Analysis for Insight
- Aggegrate for Reporting
- Support business experiments
- Improve MTTIT - Mean Time to Idea Tested

DATA CORE: TEAM GOAL

Scalability of 100x with current team size

WHAT DOES THAT MEAN ?

- Our Data Warehouse can't handle 100x of raw data
  - Stop loading raw data into DW!
  - Generate aggregates and roll-ups outside of DW
  - Load converted data to DW
- Provide some means of access to raw data for ad-hoc analysis

BUILD A SELF-SERVICE STREAMING DATA PLATFORM TO SUPPORT 20+ DEV TEAMS AT 100X CURRENT USAGE

BUILD A SELF-SERVICE ~~STREAMING~~ DATA PLATFORM TO SUPPORT 20+ DEV TEAMS AT 100X CURRENT USAGE

BUILD A ~~SELF-SERVICE STREAMING~~ DATA PLATFORM TO SUPPORT 20+ DEV TEAMS AT 100X CURRENT USAGE

BUILD A ~~SELF-SERVICE STREAMING~~ DATA PLATFORM TO SUPPORT *ONE* DEV TEAM AT 100X CURRENT USAGE

BUILD A ~~SELF-SERVICE~~ ~~STREAMING~~ DATA PLATFORM TO SUPPORT *ONE DATA* TEAM AT 100X CURRENT USAGE

DEFINITIONS:

- OLTP vs OLAP
- Data Warehouse
- ETL
- Data Lake
- Technologies

OLTP VS OLAP

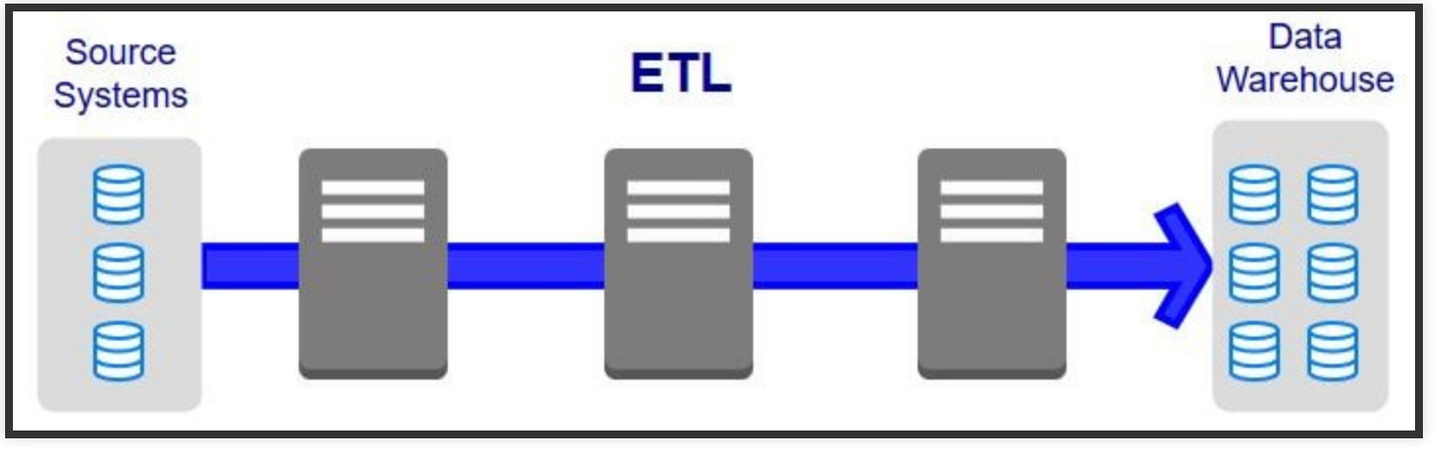
- On-Line Transaction Processing
  - Primary Database
  - High number of transactions
  - Insert / Update / Delete
  - Powers primary application
- On-Line Analytical Processing
  - Aggregated data
  - Complex Queries
  - Long running queries
  - Powers reporting

DATA WAREHOUSE

An OLAP database used for reporting.

ETL

Extract - Transform - Load



ETL

Extract

pull data from sources and convert to standardized format

Transform

apply business rules and clean

Load

Load into data warehouse for reporting

- Generate roll-ups with SQL
- Run regular and ad-hoc reports against roll-ups

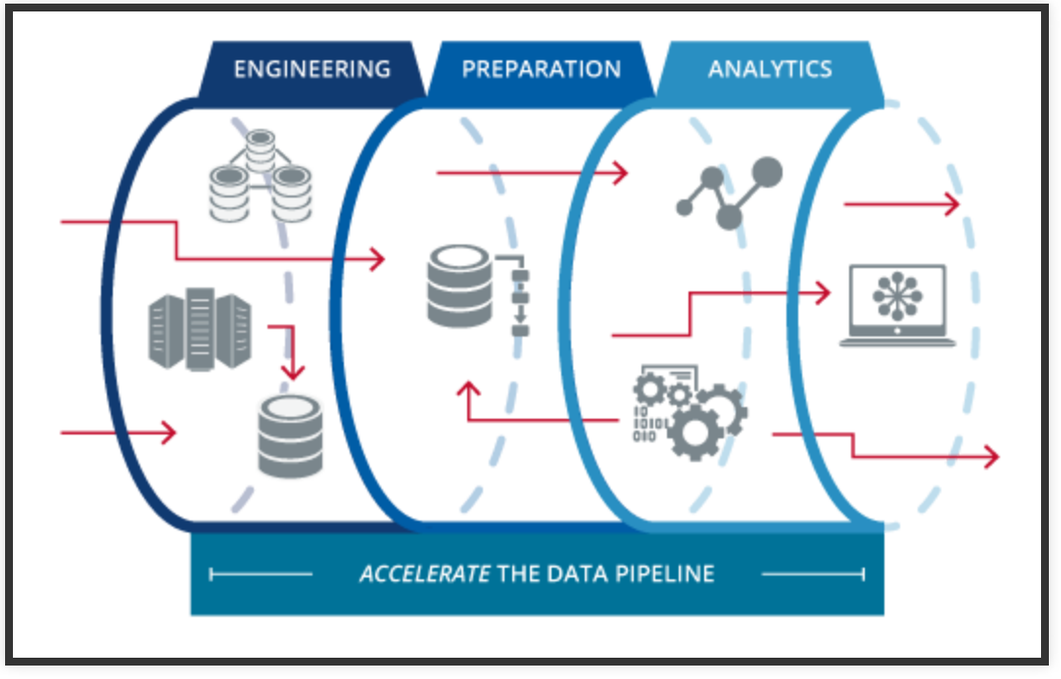
DATA LAKE

Artisanal and Unfiltered

DATA WAREHOUSE	vs.	DATA LAKE
structured, processed	**DATA**	structured / semi-structured / unstructured, raw
schema-on-write	**PROCESSING**	schema-on-read
expensive for large data volumes	**STORAGE**	designed for low-cost storage
less agile, fixed configuration	**AGILITY**	highly agile, configure and reconfigure as needed
mature	**SECURITY**	maturing
business professionals	**USERS**	data scientists et. al.

DATA PIPELINE

System to collect, clean, query, aggregate and publish data from source to analytic consumer.



TECHNOLOGIES

Mapping of Open Source and Amazon products

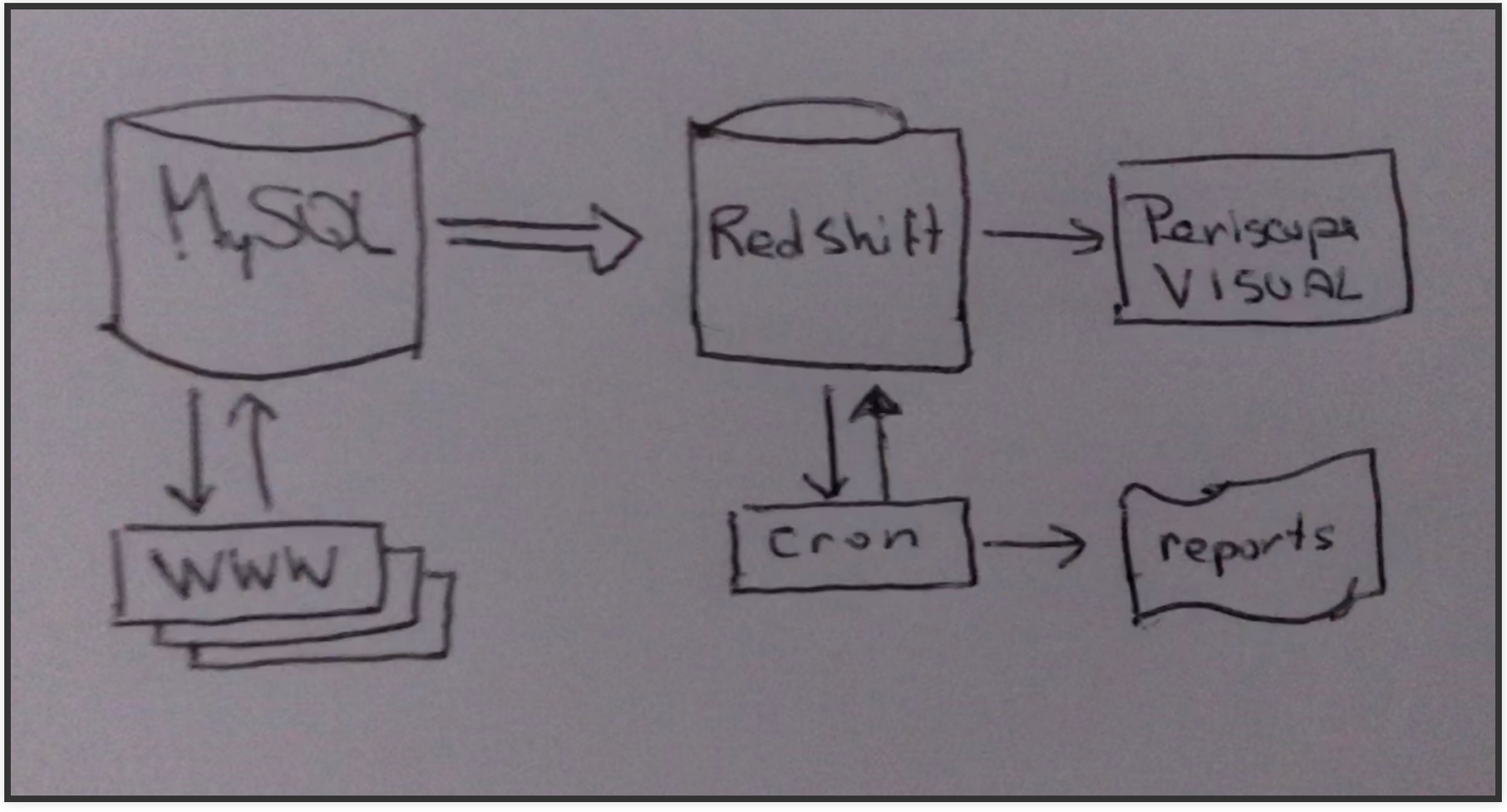
Open Source	Amazon Product	Description
Presto	Athena	Distributed SQL Query Engine
	Redshift Spectrum	Run Redshift queries against S3
Kafka	Kinesis	Distributed Streaming Platform
Ceph, etc	S3	Extensible object/file store

TECHNOLOGIES [CONT]

Open Source	Amazon Product	Description
Apache Spark		Engine for large-scale data processing.
Apache Spark SQL		module for working with structured data.
Apache Airflow		workflow platform
Hive Metastore	AWS Glue Data Catalog	central repository to store structural and operational metadata for all your data assets

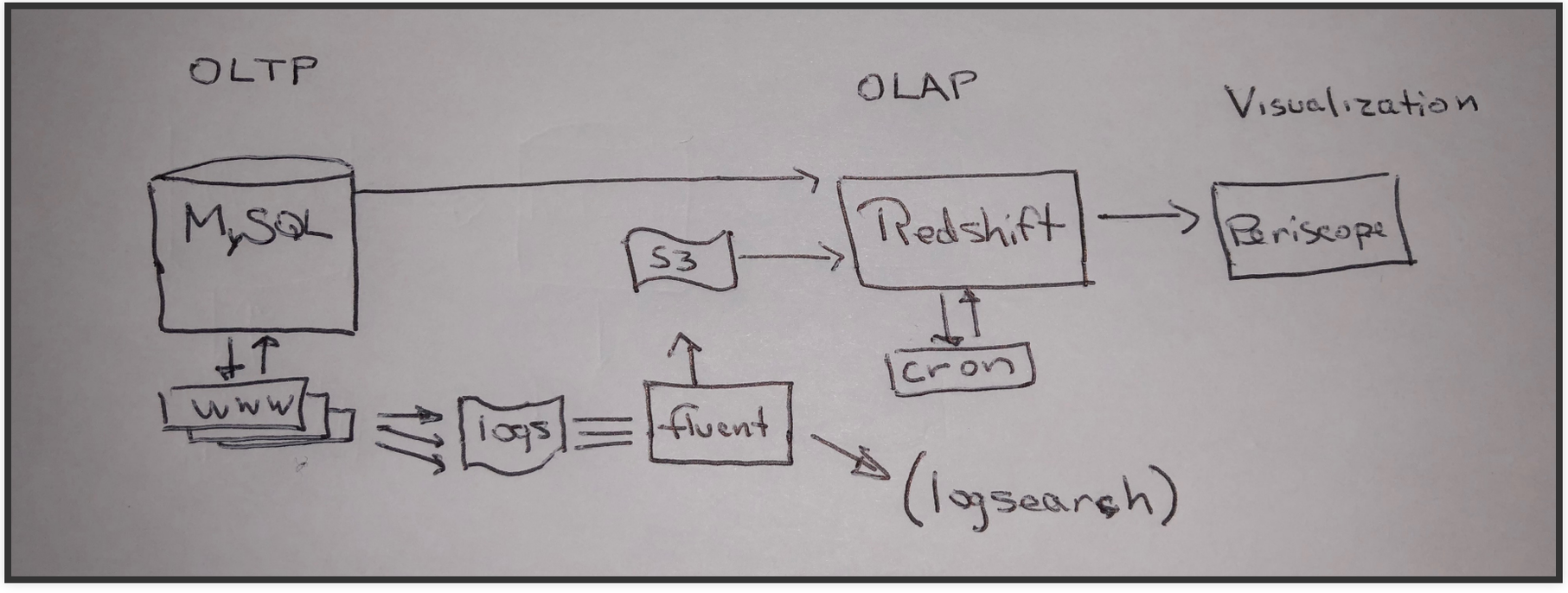
SYSTEM DIAGRAMS

WHERE WE WERE



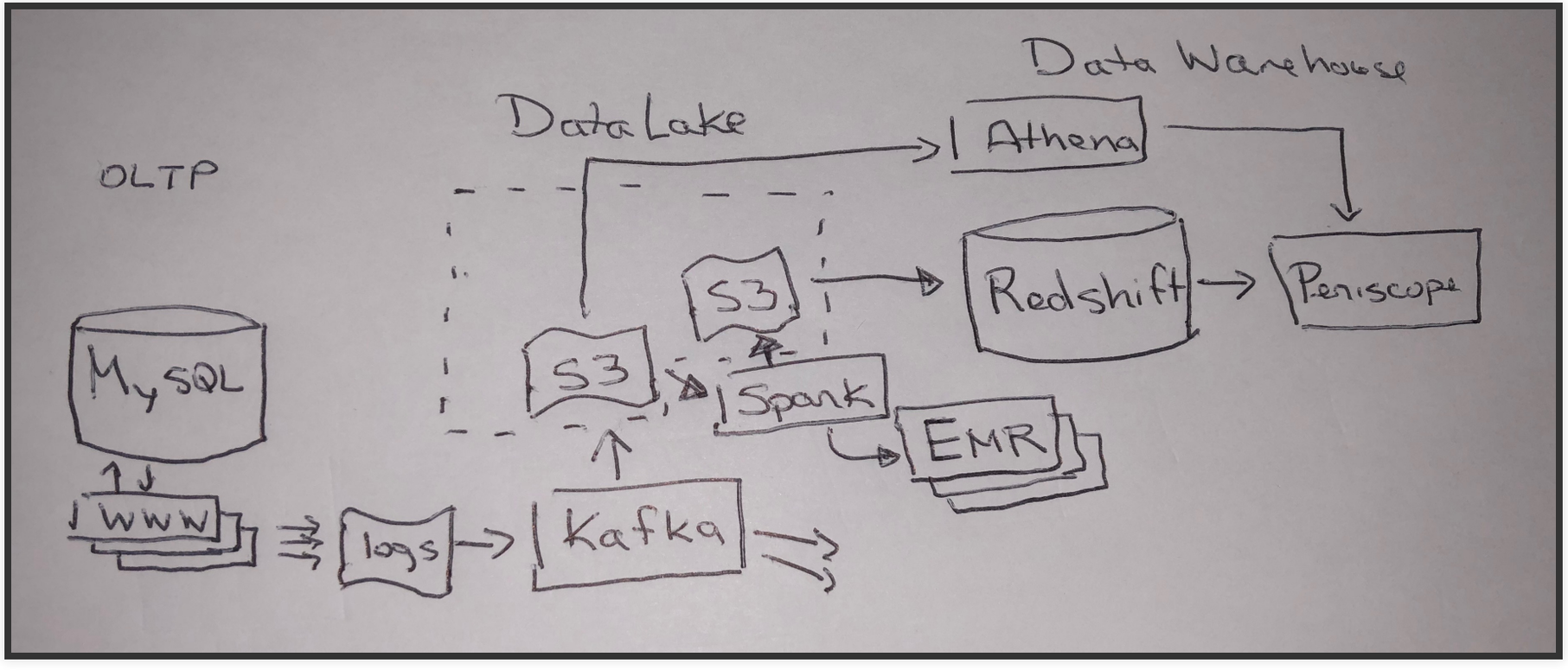
Scaling limit: Raw log data overwhelmed primary database

WHERE WE ARE



Scaling limit: Raw log data overwhelmed primary database + Cron management

WHERE WE'RE GOING



Scaling limit: ?

DATA PIPELINE PIECES

- Ingres
- Storage
- Data Formats
- Queries
- Meta data
- Workflow
- Scheduling

INGRES: JSON + KAFKA

- All logs generated as json lines
- Logs are tailed and published to Kafka
- Hourly Buckets by type from Kafka stream

STORAGE: S3

- Raw logs are source-of-truth for system
- Raw logs are stored in a production-logs bucket
- Structured path `datatype/yyyy/mm/dd/...`
- Encrypted S3 bucket

DATA FORMATS

- json lines for all logs
- Apache parquet for derived sources

QUERIES: SPARK AND SPARKSQL

- Spark: Scala and python
  - spark build tool: (sbt vs maven vs gradle)?
- SparkSQL: SQL queries over JDBC
- Athena: Interactive queries from AWS console

METASTORE: AWS GLUE DATA CATALOG

- hive-compatible metadata
- Works across SparkSQL, Athena and Spectrum
- Daily partition metadata added by cron job.

WORKFLOW

Luigi vs Azkaban vs Oozie vs Airflow

- Luigi
  - python, Spotify, code-based DAG
- Azkaban
  - java, LinkedIn, GUI, hadoop only, time-based scheduling
- Oozie
  - worklow scheduler for hadoop
- Airflow
  - python, AirBNB, code-based + GUI,

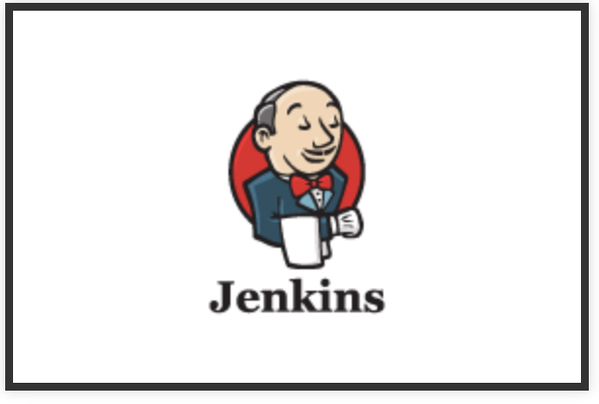
https://www.bizety.com/2017/06/05/open-source-data-pipeline-luigi-vs-azkaban-vs-oozie-vs-airflow/

AIRFLOW


Apache Airflow (incubating)
Use Apache Airflow (incubating) to author workflows as directed acyclic graphs (DAGs) of tasks
6,015
1,739

SCHEDULING - JENKINS

- Automatic and manual triggers
- Compile sources for a workflow into binary artifacts
- Launch transient EMR cluster to run job



EGRESS

- Export aggregated data to redshift as Data Warehouse
- Query Redshift via Periscope for Visualization and Dashboards

STATUS

QUESTIONS?

CONTACT:

Twitter	@spazm
Github	spazm
Email	andrew.grangaard+scale2018@gmail.com
Blog	spaz.rocks

https://spazm.github.io/slides/building_a_data_pipeline-scale16x/

EXTRA

WHAT?

- aggregation of raw data
- storage of raw data
- cleaning of data
-

SUMMARY

I'll share the architecture we design based on the trade-offs we considered and the choices we've made.

Building a data pipeline for stats and analysis is a big job. We have a cornucopia of open source tools to choose from and so many decisions to make regarding:

Tools orchestration storage formats streaming compute SQL integration data ingress, egress job vetting data integrity