STYLIGHT

Building data pipelines: from simple to more advanced - hands-on experience

SERGII KHOMENKO, DATA SCIENTIST,
SERGII.KHOMENKO@STYLIGHT.COM, @lcod3r

STYLIGHT.COM

AGENDA

Who? What? Why?
The Good, The Bad And The Legacy
Open Source stack
Amazon AWS
Google Cloud
Tips, tricks and best practices

STYLIGHT

In computing, a pipeline is a set of data processing elements connected in series, where the output of one element is the input of the next one.

Sergii Khomenko

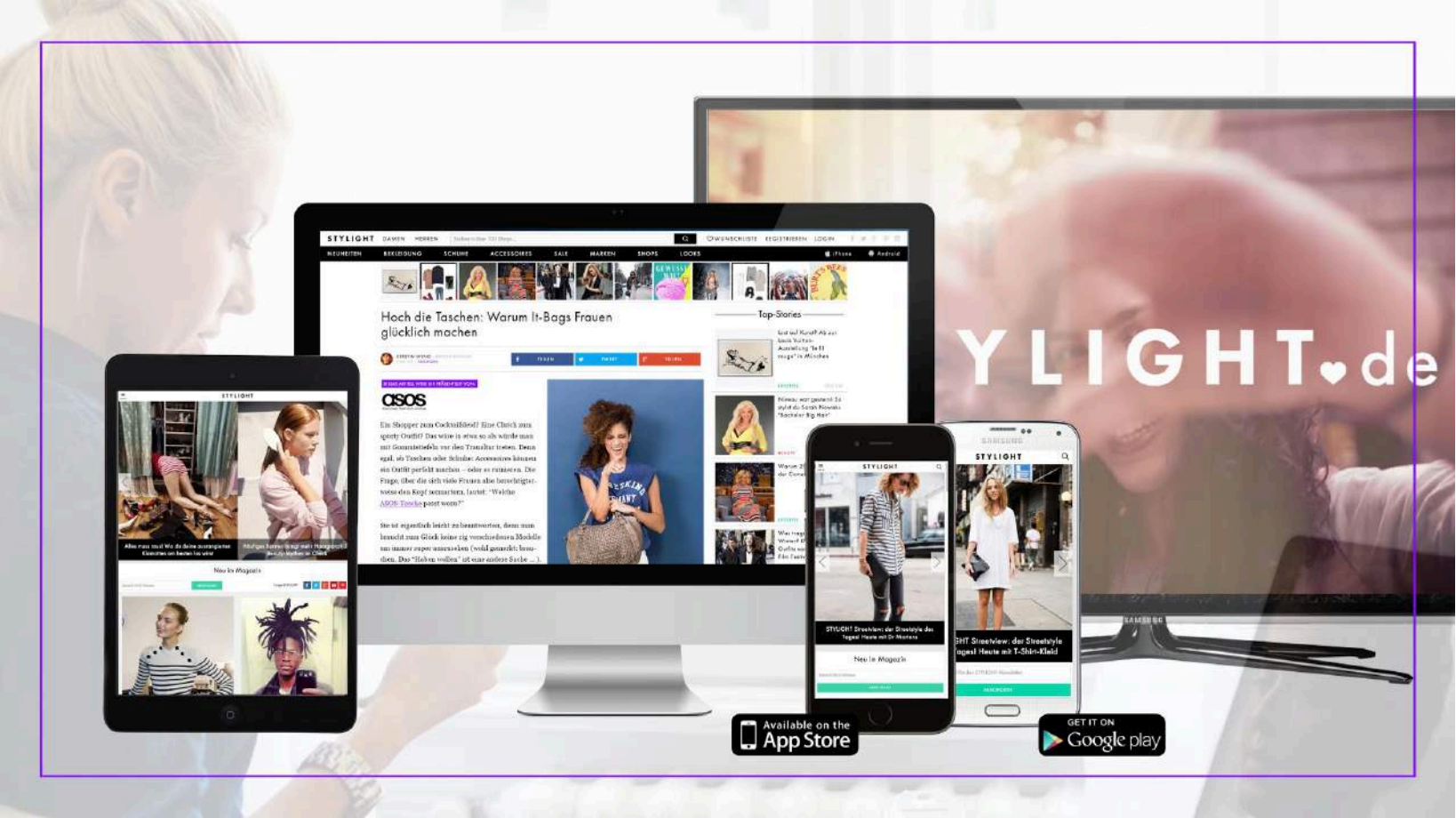
Data Scientist at one of the biggest fashion communities, STYLIGHT. Data analysis and visualization hobbyist.

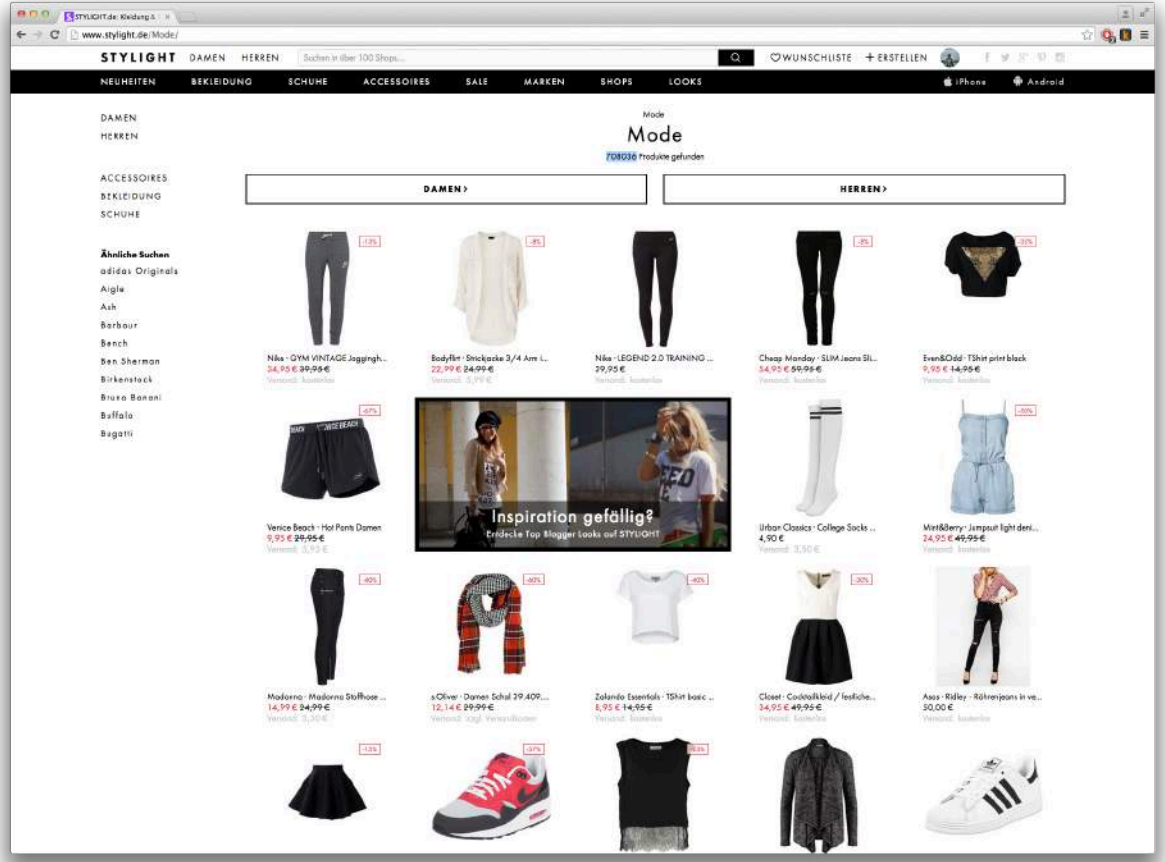
Speaker at Berlin Buzzwords 2014, ApacheCon Europe 2014, Puppet Camp London 2015

Founder and speaker at Munich Golang UG, Munich Tableau UG.

Speaker at Munich UseR Group, Munich Search UG, Munich Quantified Self UG, Munich DataGeeks.

STYLIGHT

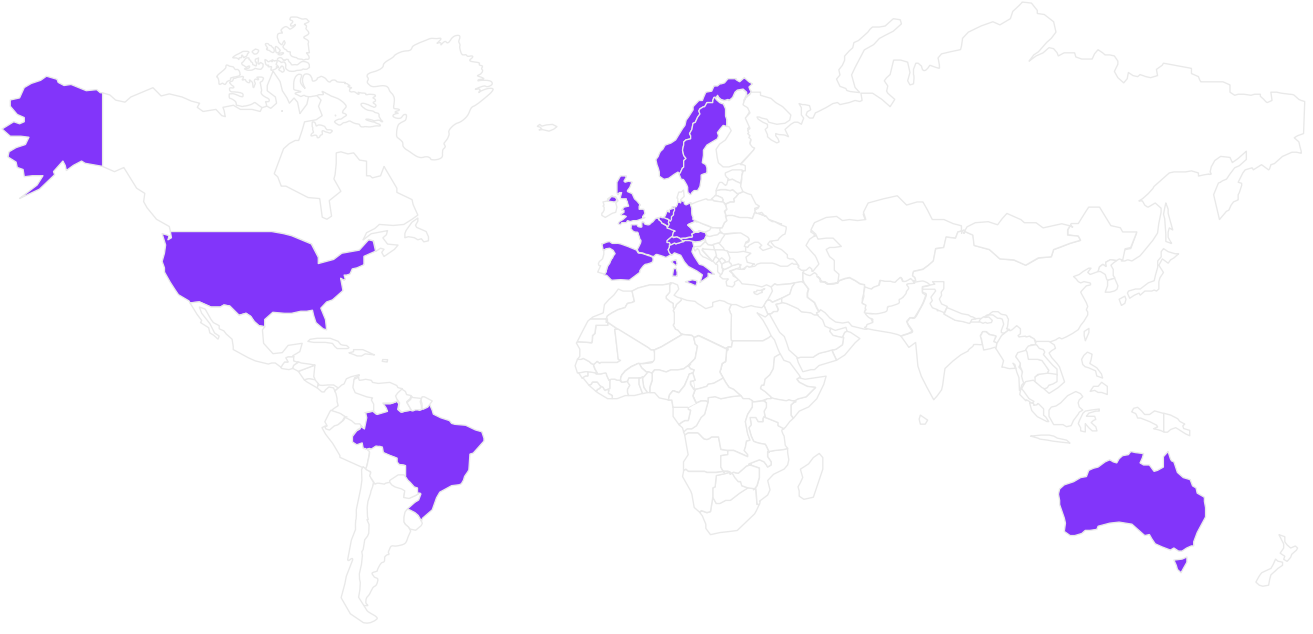

STYLIGHT
Hoch die Taschen: Warum It-Bags Frauen glücklich machen
Top-Stories
asos
Neu im Magazin
STYLIGHT Streetview: der Streetstyle
Available on the App Store
GET IT ON Google play


STYLIGHT.de: Kleidung &
www.stylight.de/Mode/
STYLIGHT
DAMEN
HERREN
Suchen in über 100 Shops...
WUNSCHLISTE
+ ERSTELLEN
NEUHEITEN
BEKLEIDUNG
SCHUHE
ACCESSOIRES
SALE
MARKEN
SHOPS
LOOKS
iPhone
Android
DAMEN
HERREN
ACCESSOIRES
BEKLEIDUNG
SCHUHE
Ähnliche Suchen
adidas Originals
Aigle
Ash
Barbour
Bench
Ben Sherman
Birkenstock
Bruno Banani
Buffalo
Bugatti
Mode
Mode
70803 Produkte gefunden
DAMEN ›
HERREN ›
Nike · GYM VINTAGE Jogginghose
34,95 € 39,95 €
Bodyflirt · Strickjacke 3/4 Arm L...
22,99 € 24,99 €
Nike · LEGEND 2.0 TRAINING ...
29,95 €
Cheap Monday · SLIM Jeans Slim...
54,95 € 59,95 €
Even&Odd · TShirt print black
9,95 € 14,95 €
Venice Beach · Hot Pants Damen
9,95 € 29,95 €
Inspiration gefällig?
Entdecke Top Blogger Looks auf STYLIGHT
Urban Classics · College Socks ...
4,90 €
Mint&Berry · Jumpsuit light deni...
24,95 € 49,95 €
Madonna · Madonna Stoffhose ...
14,99 € 24,99 €
s.Oliver · Damen Schal 29.409....
12,14 € 29,99 €
Zalando Essentials · TShirt basic ...
6,95 € 14,95 €
Closet · Cocktailkleid / festliche...
34,95 € 49,95 €
Asos · Ridley · Röhrenjeans in ve...
50,00 €

STYLIGHT

STYLIGHT – international community

Live in 14 countries



STYLIGHT


STYLIGHT
The Good, The Bad And The Legacy
STYLIGHT.COM

Data sources

Where your data coming form

Sources of data:

- Web tracking
  - Metrics tracking
  - Behavior tracking
- Business intelligence ETL
- Internal Services
  - ML tagging service

STYLIGHT

Access patterns

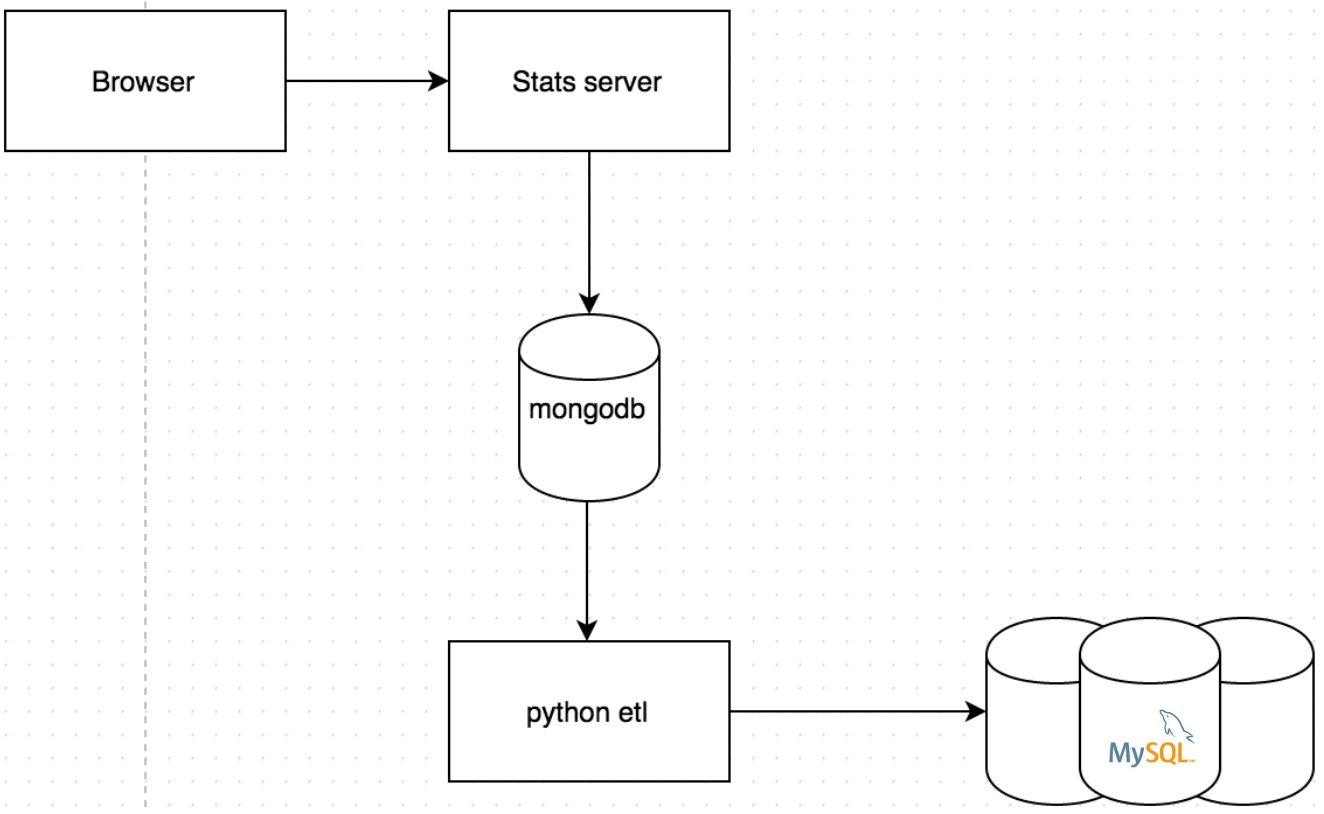
- Real-time
- Nearly real-time
- Daily batches

STYLIGHT

python

Google Analytics

STYLIGHT

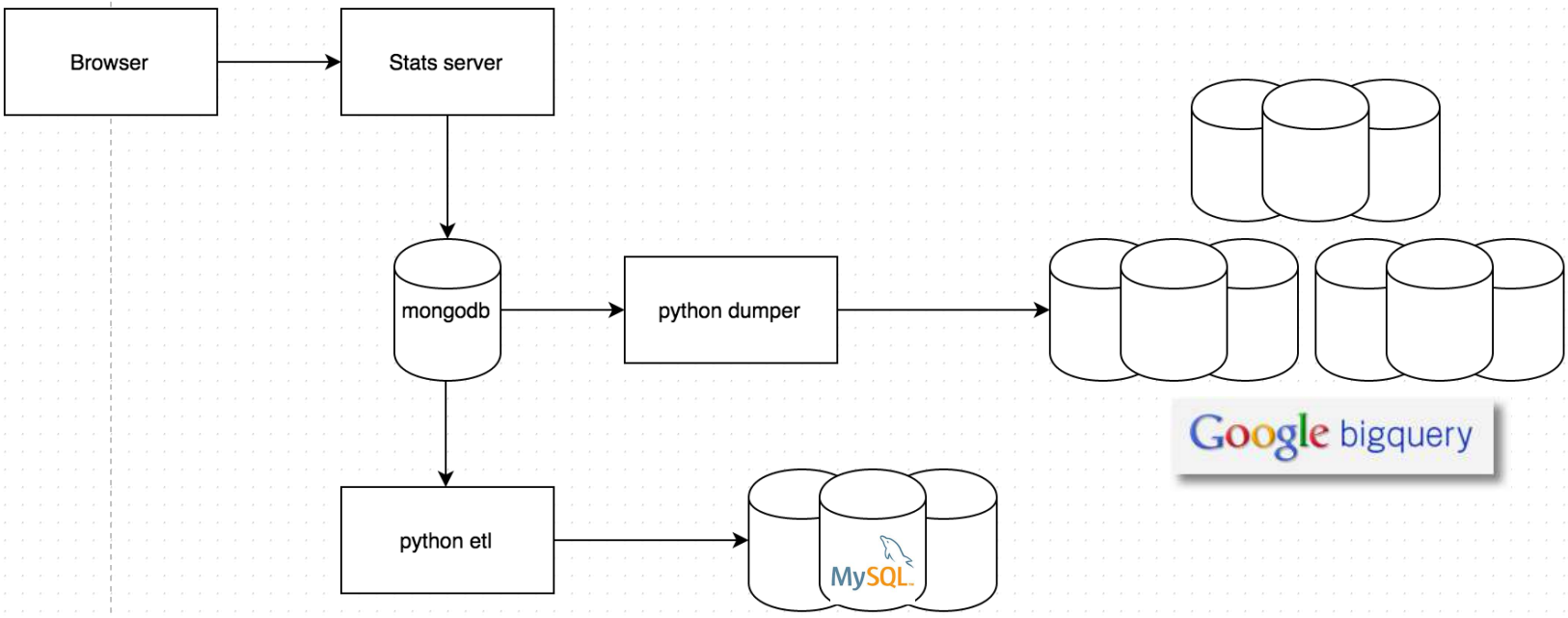

Browser
Stats server
mongodb
python etl
MySQL

Properties

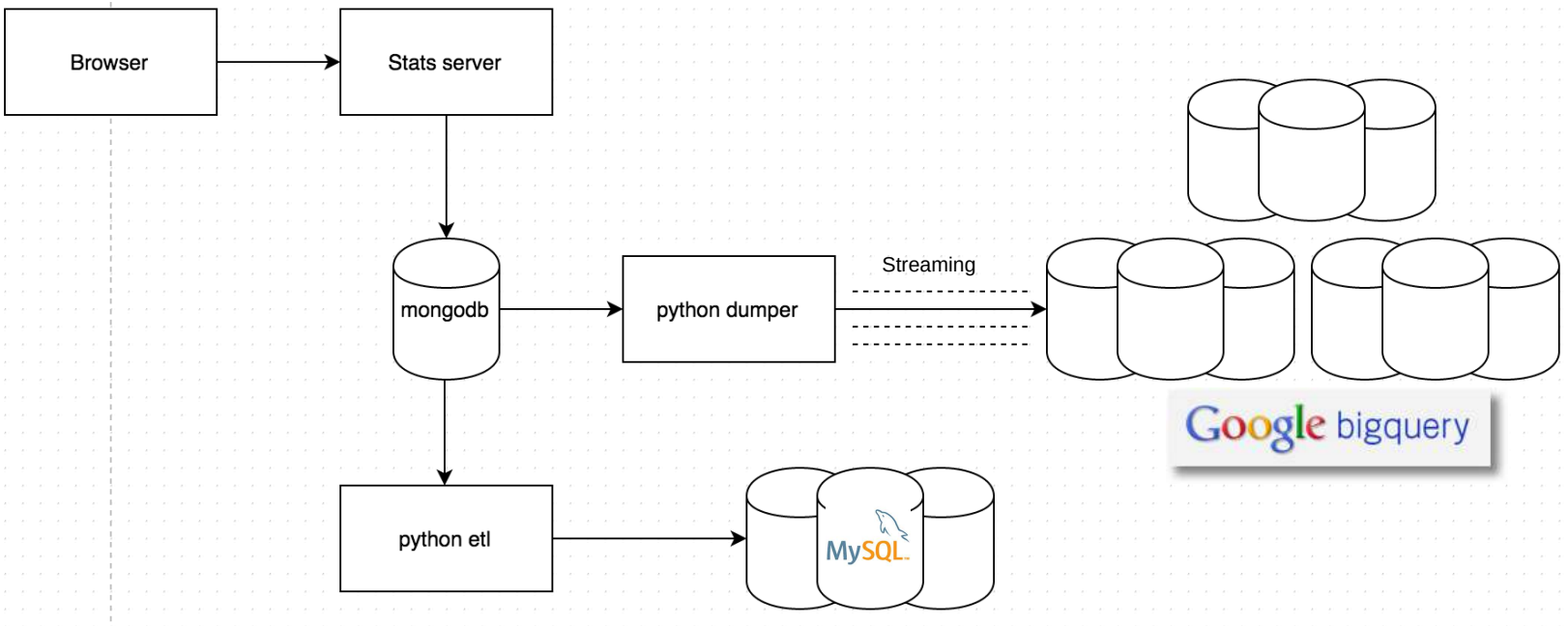
- Data consistency
- Doesn’t scale
- Hard to add new sources
- Complex system
- Many interfaces

- As lean and legacy as possible
- No need for special services

STYLIGHT

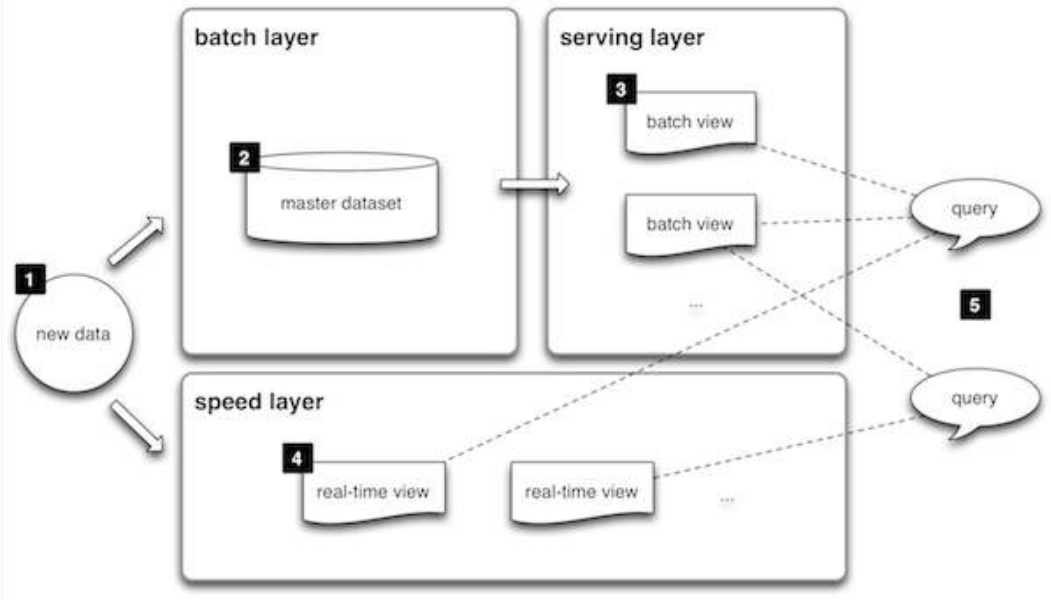

Browser
Stats server
mongodb
python dumper
Google bigquery
python etl
MySQL

STYLIGHT


Browser
Stats server
mongodb
python dumper
Streaming
Google bigquery
python etl
MySQL

Open Source Stack

STYLIGHT.COM

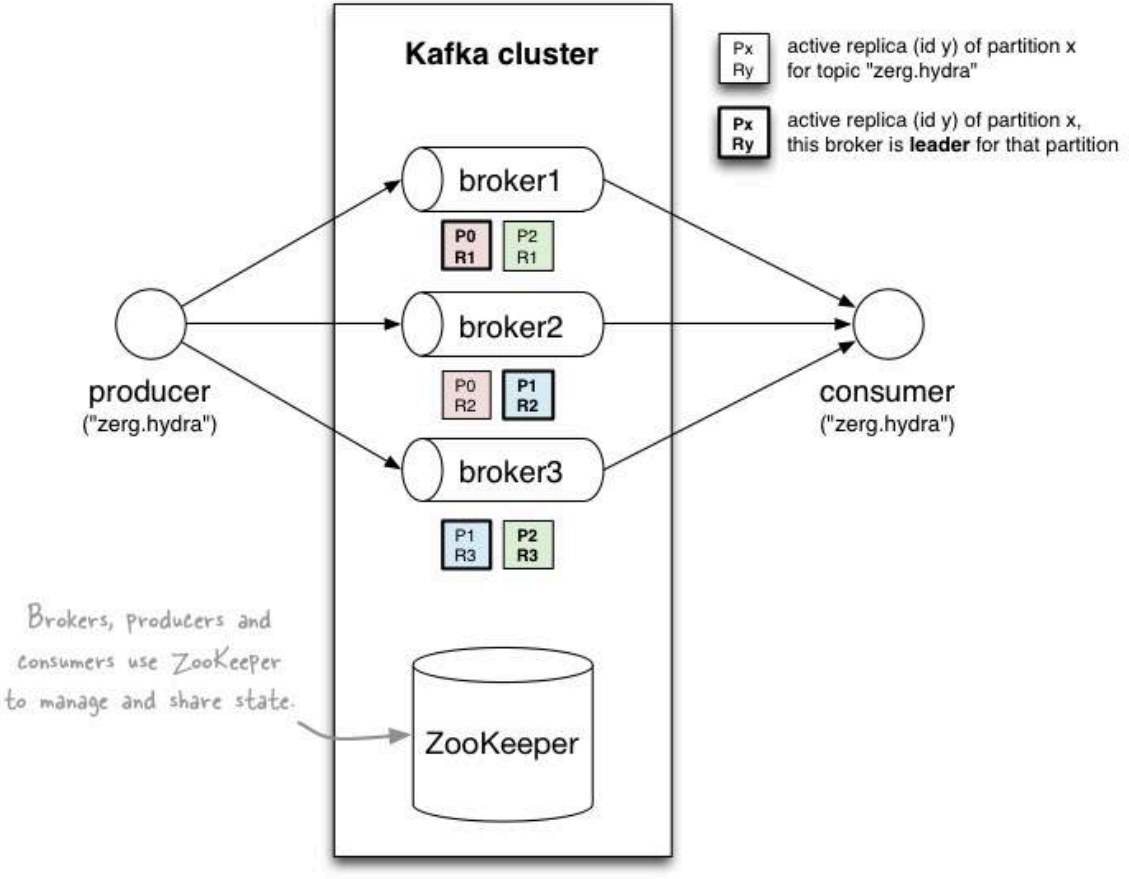


http://lambda-architecture.net/

STYLIGHT

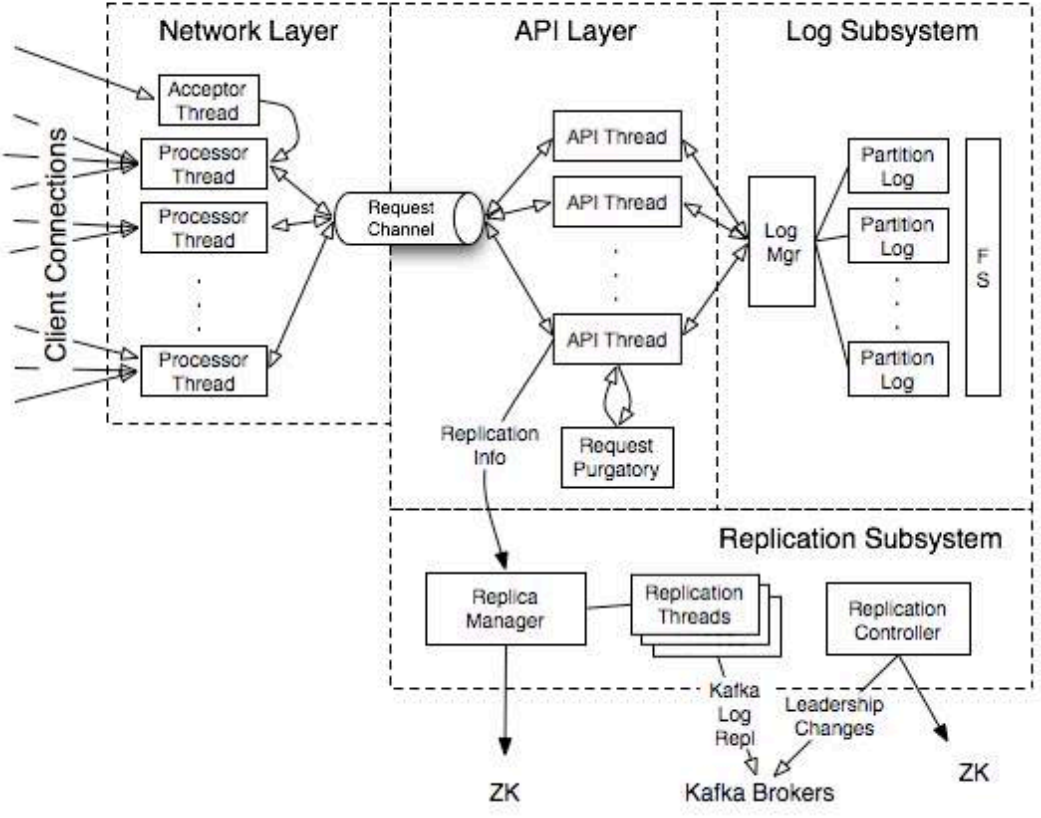
Apache Kafka is publish-subscribe messaging rethought as a distributed commit log.

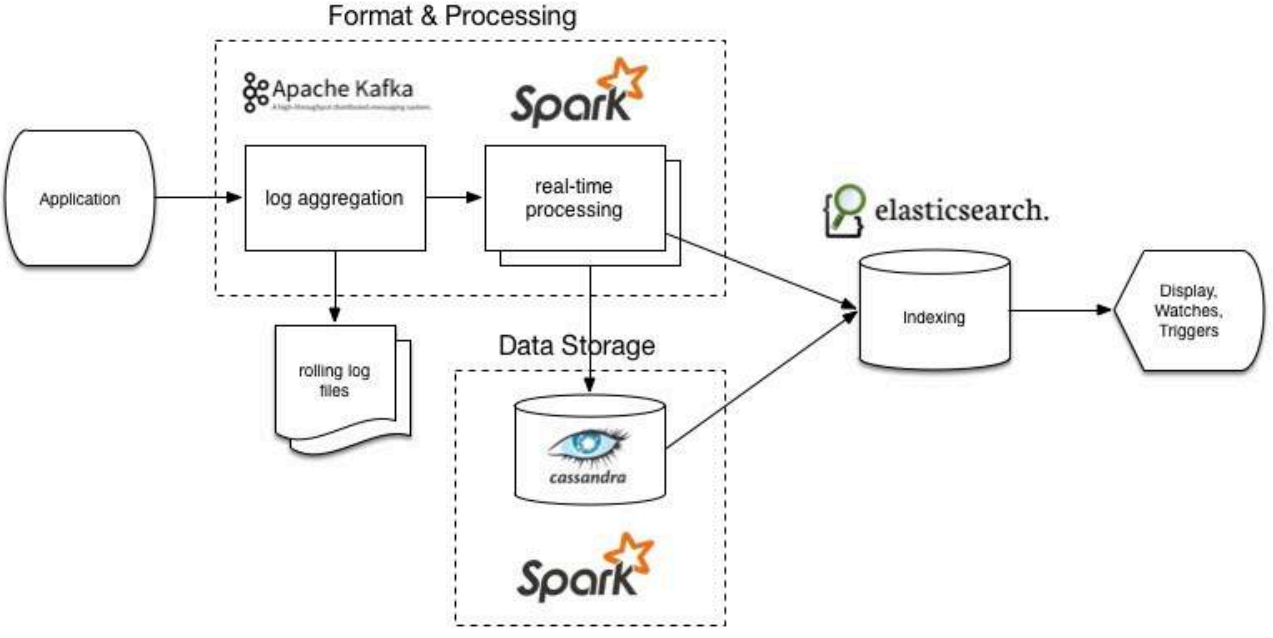
STYLIGHT


Kafka cluster
Px Ry
active replica (id y) of partition x for topic "zerg.hydra"
Px Ry
active replica (id y) of partition x, this broker is leader for that partition
broker1
P0 R1
P2 R1
broker2
P0 R2
P1 R2
broker3
P1 R3
P2 R3
producer ("zerg.hydra")
consumer ("zerg.hydra")
Brokers, producers and consumers use ZooKeeper to manage and share state.
ZooKeeper

STYLIGHT

Kafka Broker Internals





http://www.ipponusa.com/wp-content/uploads/2014/10/spark-architecture.jpg

STYLIGHT

Results

- Scalable
- Flexible

- High costs of maintenance
- Not so easy to setup

STYLIGHT

Amazon AWS

STYLIGHT.COM

Amazon Web Services

Compute & Networking

- **Direct Connect** — Dedicated Network Connection to AWS
- **EC2** — Virtual Servers in the Cloud
- **Route 53** — Scalable Domain Name System
- **VPC** — Isolated Cloud Resources

Storage & Content Delivery

- **CloudFront** — Global Content Delivery Network
- **Glacier** — Archive Storage in the Cloud
- **S3** — Scalable Storage in the Cloud
- **Storage Gateway** — Integrates On-Premises IT Environments with Cloud Storage

Database

- **DynamoDB** — Predictable and Scalable NoSQL Data Store
- **ElastiCache** — In-Memory Cache
- **RDS** — Managed Relational Database Service
- **Redshift** — Managed Petabyte-Scale Data Warehouse Service

Deployment & Management

- **CloudFormation** — Templated AWS Resource Creation
- **CloudTrail** — User Activity and Change Tracking
- **CloudWatch** — Resource and Application Monitoring
- **Elastic Beanstalk** — AWS Application Container
- **IAM** — Secure AWS Access Control
- **OpsWorks** — DevOps Application Management Service
- **Trusted Advisor** — AWS Cloud Optimization Expert

Analytics

- **Data Pipeline** — Orchestration for Data-Driven Workflows
- **Elastic MapReduce** — Managed Hadoop Framework
- **Kinesis** — Real-time Processing of Streaming Big Data

Mobile Services

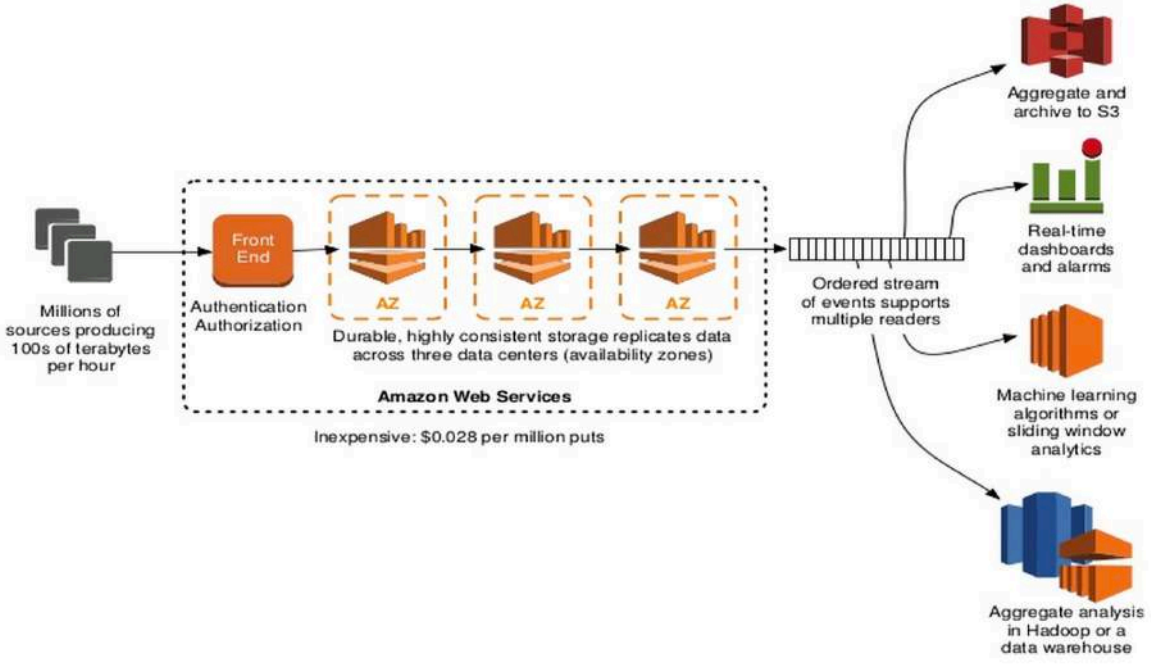
- **Cognito** — User Identity and App Data Synchronization
- **Mobile Analytics** — Understand App Usage Data at Scale
- **SNS** — Push Notification Service

App Services

- **AppStream** — Low Latency Application Streaming
- **CloudSearch** — Managed Search Service
- **Elastic Transcoder** — Easy-to-use Scalable Media Transcoding
- **SES** — Email Sending Service
- **SQS** — Message Queue Service
- **SWF** — Workflow Service for Coordinating Application Components

Applications

- **WorkSpaces** — Desktops in the Cloud
- **Zocalo** — Secure Enterprise Storage and Sharing Service


Front End
Authentication
Authorization
AZ
AZ
AZ
Durable, highly consistent storage replicates data across three data centers (availability zones)
Amazon Web Services
Millions of sources producing 100s of terabytes per hour
Inexpensive: $0.028 per million puts
Ordered stream of events supports multiple readers
Aggregate and archive to S3
Real-time dashboards and alarms
Machine learning algorithms or sliding window analytics
Aggregate analysis in Hadoop or a data warehouse

STYLIGHT

Real-time Ingest

Highly Scalable

Durable

Elastic

Replay-able Reads

Continuous Processing FX

Elastic

Load-balancing incoming streams

Fault-tolerance, Checkpoint / Replay

Enable multiple processing apps in paralle

Managed Service

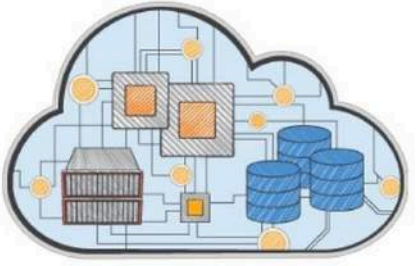
Low end-to-end latency

Enable data movement into Stores/ Processing Engines

STYLIGHT

AWS Lambda: A compute service that runs your code in response to events

Lambda functions: Stateless, event-driven code execution



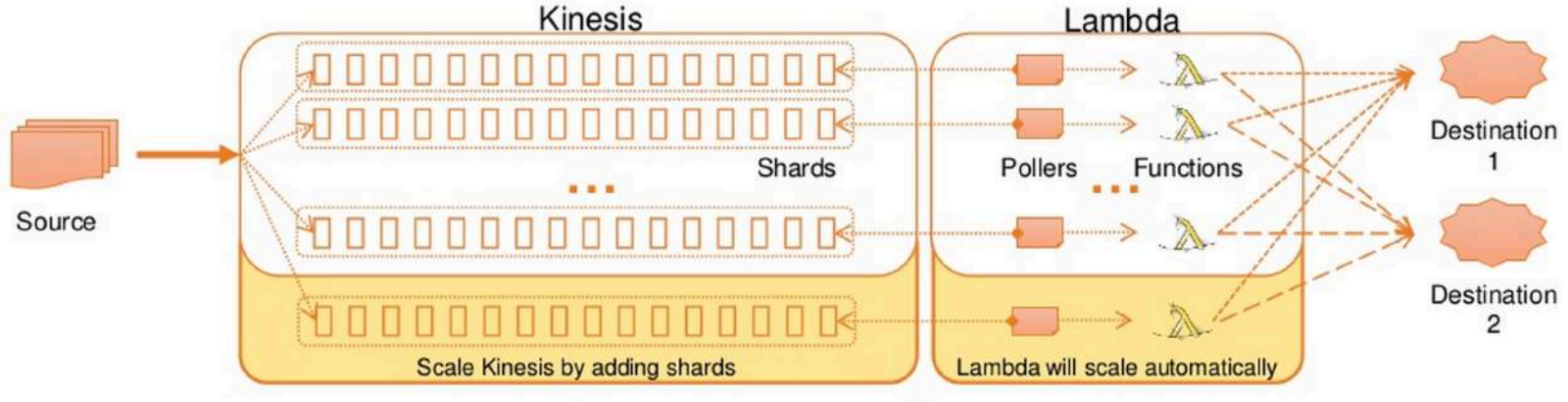
Triggered by events:

- Put to an Amazon S3 bucket
- Record in an Amazon Kinesis stream
- Direct sync and async invocations

Makes it easy to

- Build back-end services that perform at scale
- Perform data-driven auditing, analysis, and notification

STYLIGHT

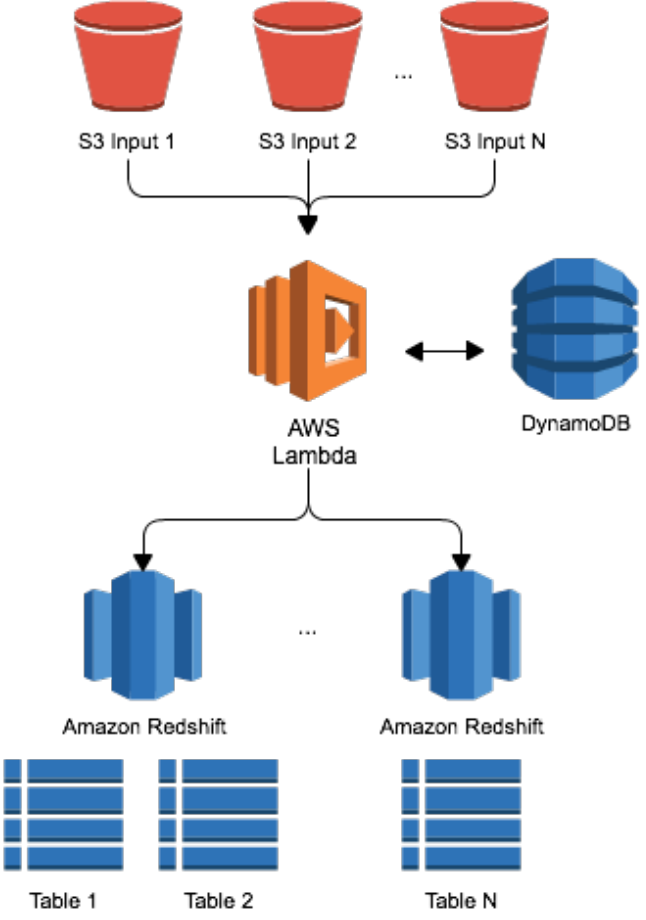

Kinesis
Lambda
Source
Shards
Pollers
Functions
Destination 1
Destination 2
Scale Kinesis by adding shards
Lambda will scale automatically

STYLIGHT

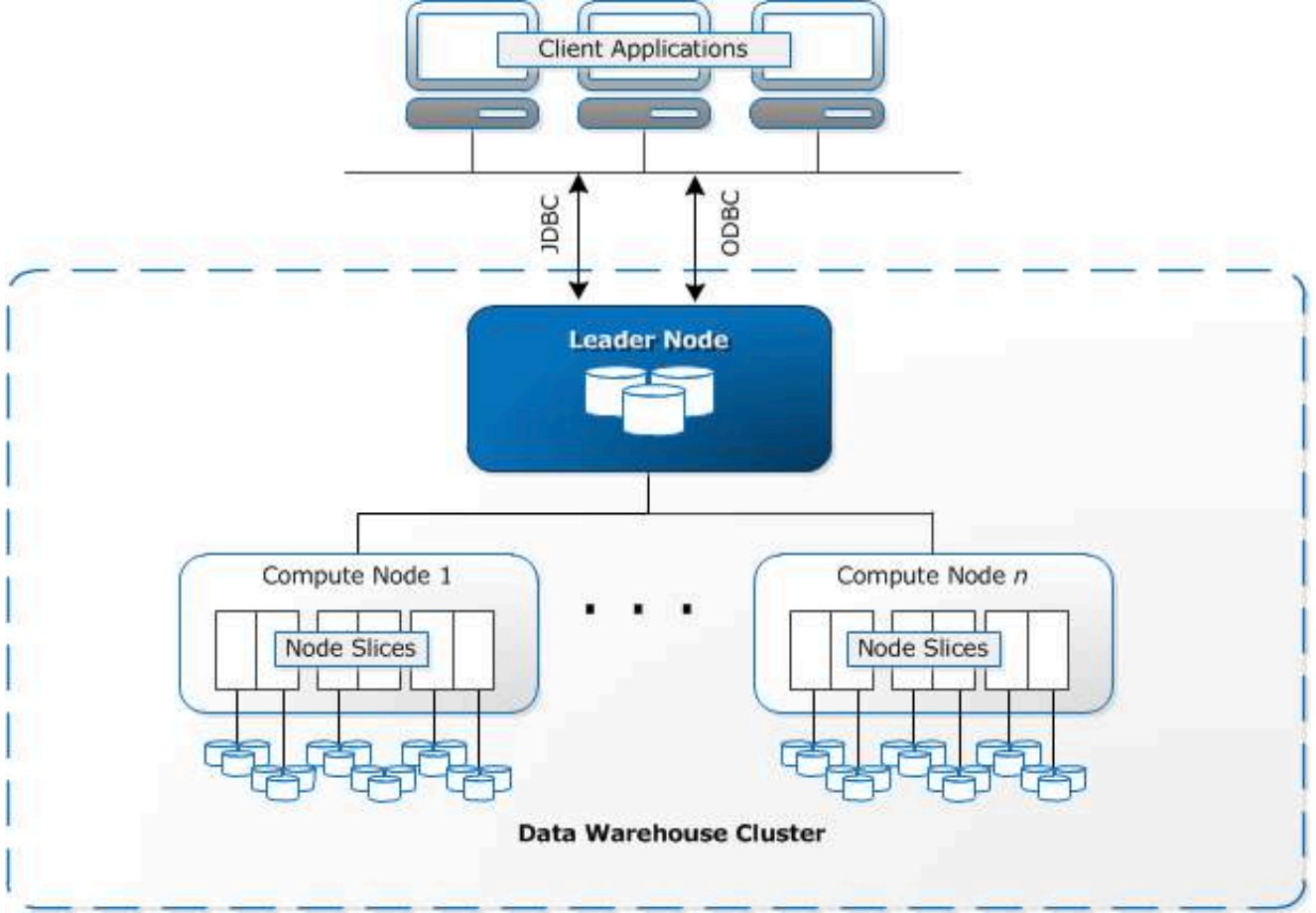
Performance tuning Kinesis as an event source

- **Batch size**: Number of records that AWS Lambda will retrieve from Kinesis at the time of invoking your function
- Increasing batch size will cause fewer Lambda function invocations with more data processed per function

- **Starting Position**: The position in the stream where Lambda starts reading
- Set to “Trim Horizon” for ordered processing (FIFO)
- Set to “Latest” for reading most recent data (LIFO)

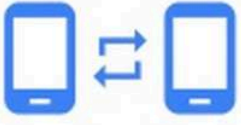

S3 Input 1
S3 Input 2
...
S3 Input N
AWS Lambda
DynamoDB
...
Amazon Redshift
Amazon Redshift
Table 1
Table 2
Table N

STYLIGHT


Client Applications
JDBC
ODBC
Leader Node
Compute Node 1
Node Slices
Compute Node n
Node Slices
Data Warehouse Cluster


STYLIGHT
Google Cloud
STYLIGHT.COM

Data on Google Cloud



Capture

Pub/Sub

Store

Storage

SQL

Datastore

Process

Dataflow

Analyze

BigQuery

Dataflow

Open Source Tools

STYLIGHT

Pub/Sub

Scalable, flexible, and globally available messaging

Dataflow

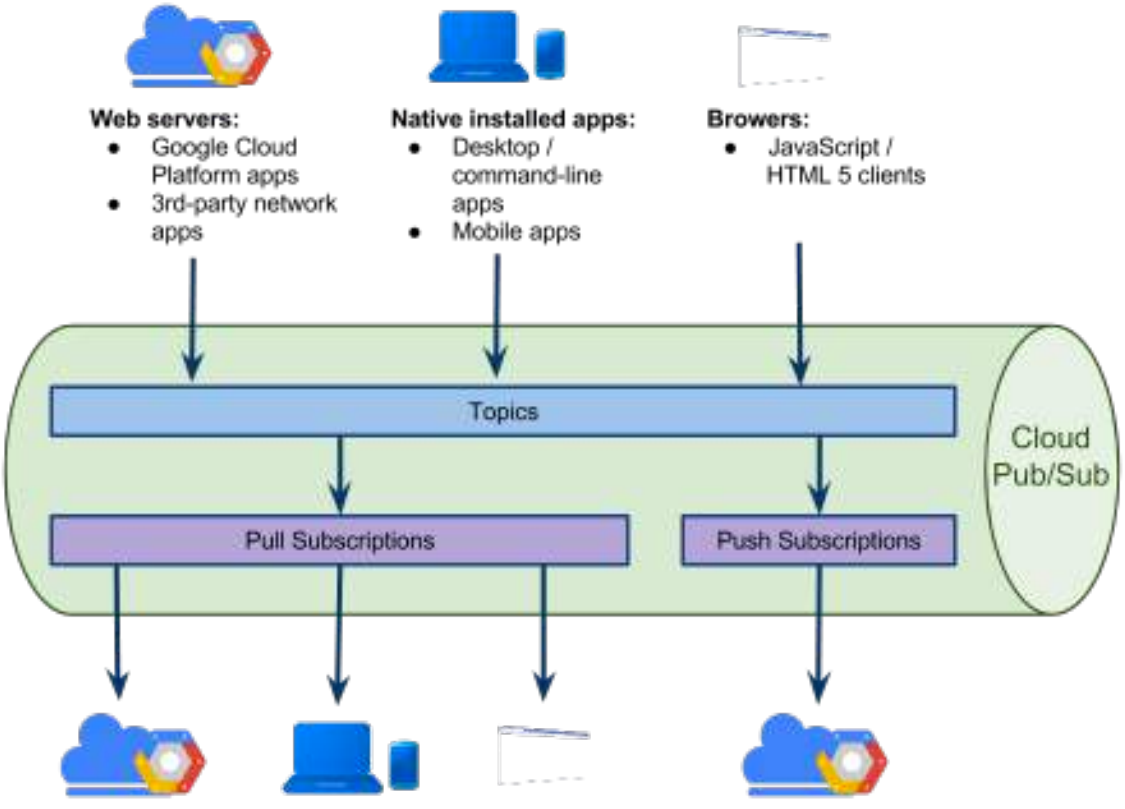
Stream & batch processing, unified and simplified

BigQuery

Ingest data at 100,000 rows per second

Fully Managed, No-Ops Services

STYLIGHT


Web servers:
Google Cloud Platform apps
3rd-party network apps
Native installed apps:
Desktop / command-line apps
Mobile apps
Browers:
JavaScript / HTML 5 clients
Topics
Pull Subscriptions
Push Subscriptions
Cloud Pub/Sub

Cloud Dataflow

Cloud Dataflow is a collection of SDKs for **building** batch or streaming parallelized data processing pipelines.

Cloud Dataflow is a fully managed service for **executing** optimized parallelized data processing pipelines.

```
Pipeline{
        Who => Inputs
        What => Transforms
        Where => Windows
        When => Watermarks + Triggers
        To => Outputs
}
```

<- At once guarantee (modulo completeness thresholds)

<- GCS, Pub/Sub, BigQuery, w/Avro, XML, JSON, ...

<- Aggregations, Filters, Joins, ...

<- Time space Fixed, Sliding, Sessions, ...

<- Correctness

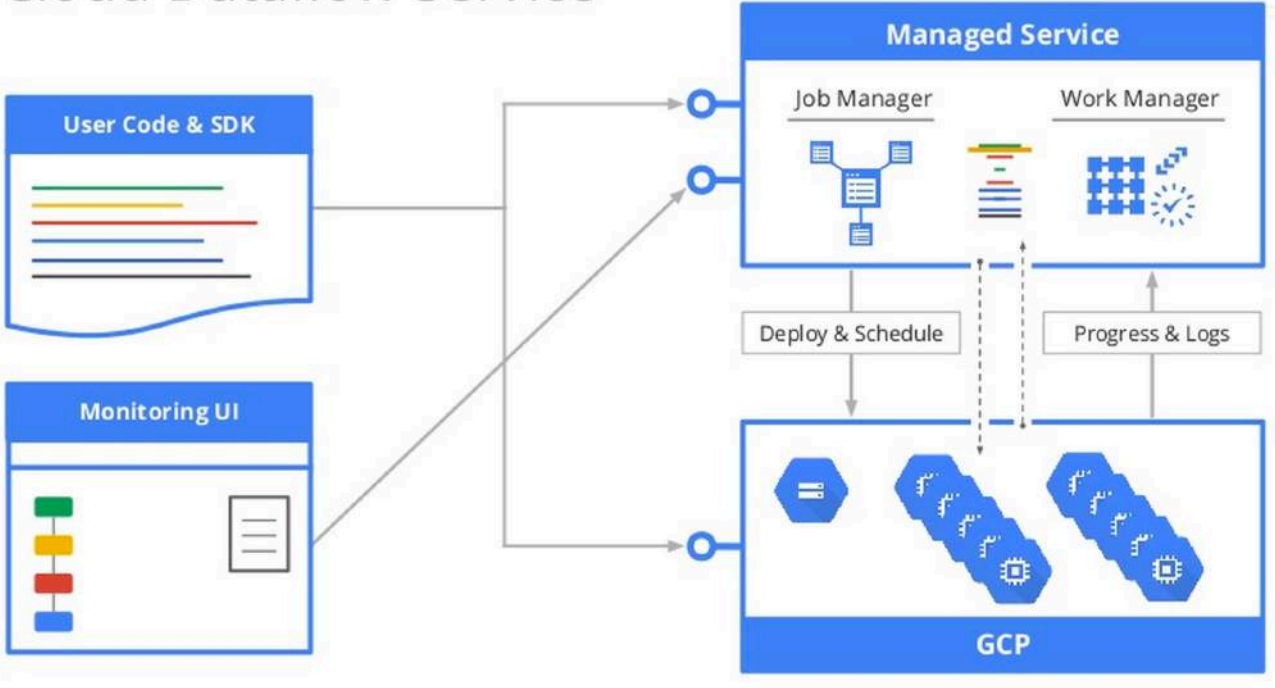
<- GCS, Pub/Sub, BigQuery, ...

STYLIGHT

Benefits of Cloud Dataflow

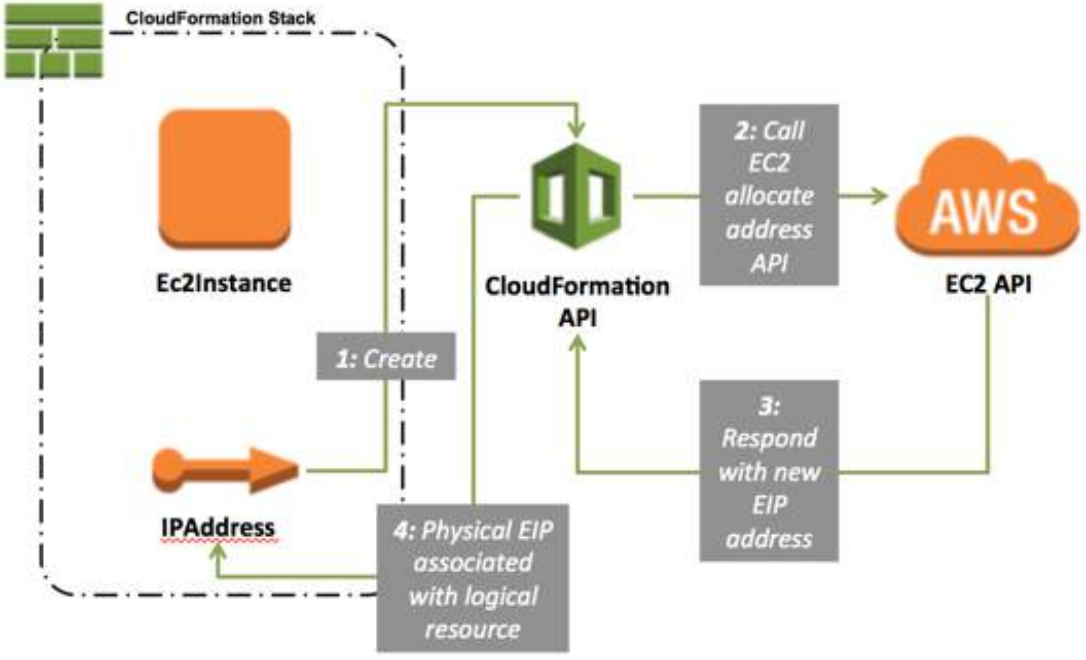
- No Ops - truly elastic data processing for the cloud
  - On demand resource allocation w/intelligent auto-scaling
  - Automated worker lifetime management
  - Automated work optimization
- Unified model - for batch & stream based processing
  - Functional programming model
  - Fine grained correctness primitives
- Open sourced SDK @ github
  - Java 7 today @ /GoogleCloudPlatform/DataflowJavaSDK
  - Python 2 in progress
  - Scala @/darkjh/scalaflow & /jhlch/scala-dataflow-dsl
  - Spark runner@ /cloudera/spark-dataflow
  - Flink runner @ /dataArtisans/flink-dataflow

Cloud Dataflow Service

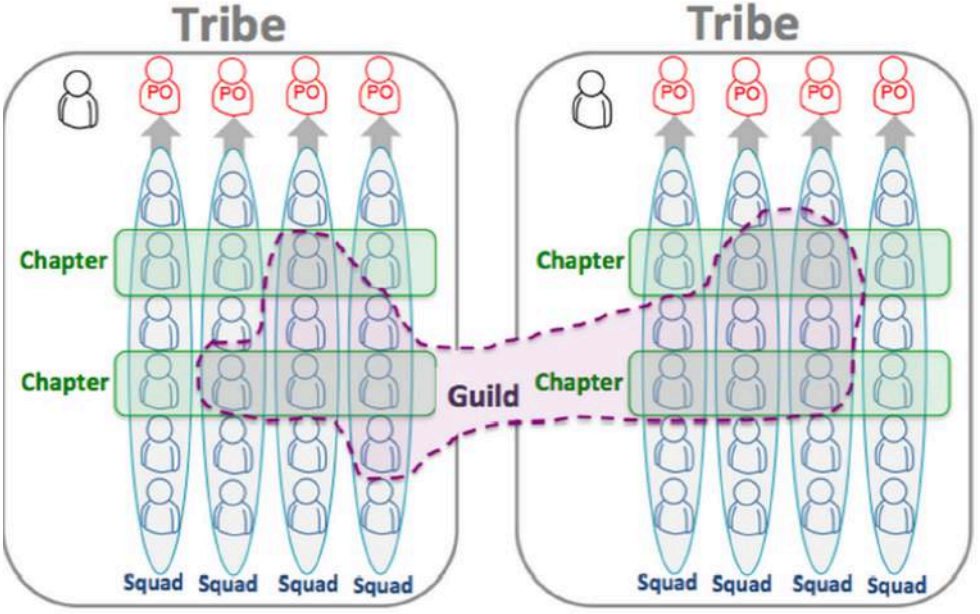



STYLIGHT
Tips, tricks and best practices
STYLIGHT.COM

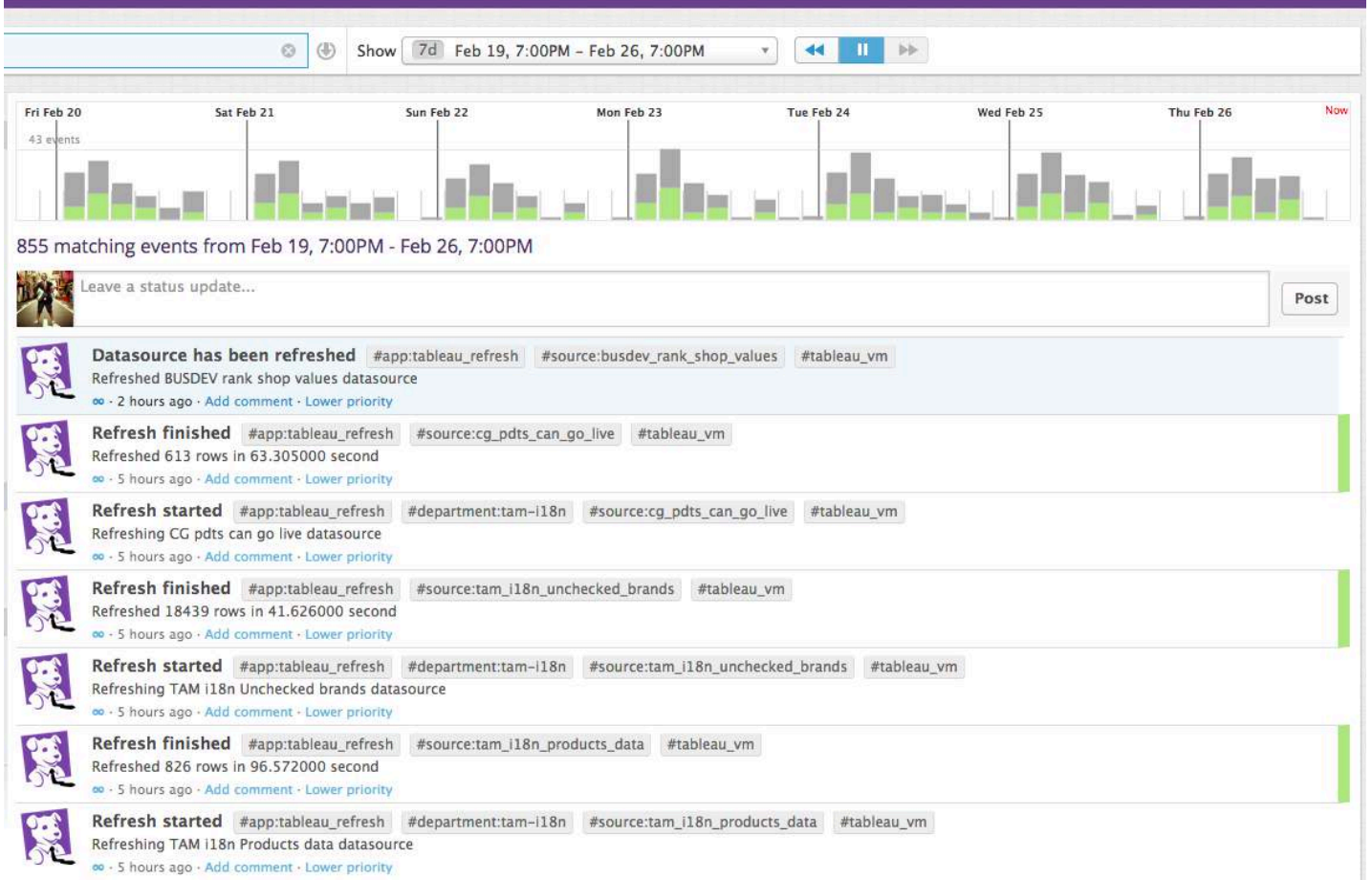
CLOUDFORMATION STACK ON AMAZON AWS



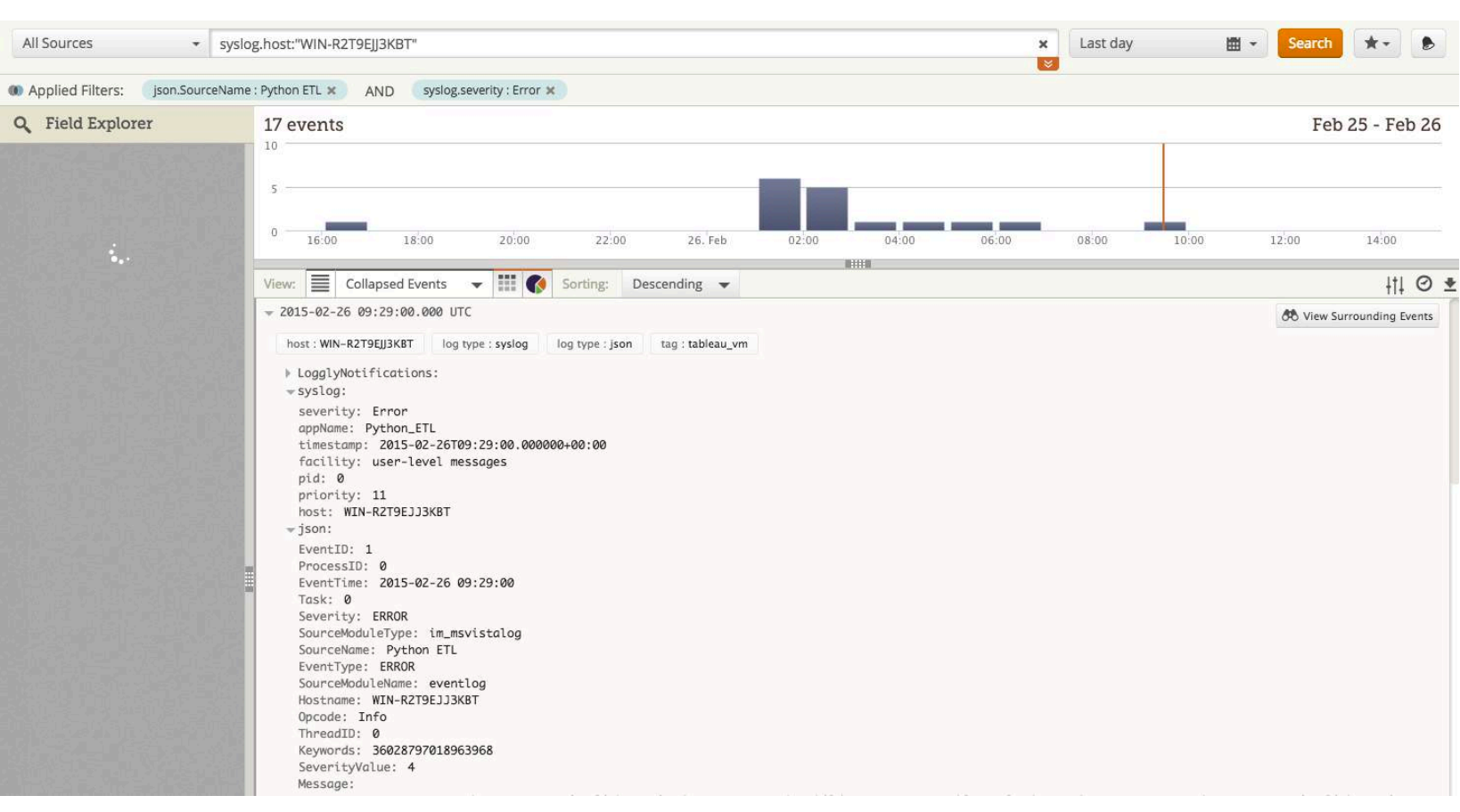
MAKE YOUR TEAM THE OWNERS OF PRODUCT AND DATA



STYLIGHT


Show
7d
Feb 19, 7:00PM – Feb 26, 7:00PM
Fri Feb 20
Sat Feb 21
Sun Feb 22
Mon Feb 23
Tue Feb 24
Wed Feb 25
Thu Feb 26
Now
43 events
855 matching events from Feb 19, 7:00PM - Feb 26, 7:00PM
Leave a status update...
Post
Datasource has been refreshed #app:tableau_refresh #source:busdev_rank_shop_values #tableau_vm
Refreshed BUSDEV rank shop values datasource
∞ · 2 hours ago · Add comment · Lower priority
Refresh finished #app:tableau_refresh #source:cg_pdts_can_go_live #tableau_vm
Refreshed 613 rows in 63.305000 second
∞ · 5 hours ago · Add comment · Lower priority
Refresh started #app:tableau_refresh #department:tam-i18n #source:cg_pdts_can_go_live #tableau_vm
Refreshing CG pdts can go live datasource
∞ · 5 hours ago · Add comment · Lower priority
Refresh finished #app:tableau_refresh #source:tam_i18n_unchecked_brands #tableau_vm
Refreshed 18439 rows in 41.626000 second
∞ · 5 hours ago · Add comment · Lower priority
Refresh started #app:tableau_refresh #department:tam-i18n #source:tam_i18n_unchecked_brands #tableau_vm
Refreshing TAM i18n Unchecked brands datasource
∞ · 5 hours ago · Add comment · Lower priority
Refresh finished #app:tableau_refresh #source:tam_i18n_products_data #tableau_vm
Refreshed 826 rows in 96.572000 second
∞ · 5 hours ago · Add comment · Lower priority
Refresh started #app:tableau_refresh #department:tam-i18n #source:tam_i18n_products_data #tableau_vm
Refreshing TAM i18n Products data datasource
∞ · 5 hours ago · Add comment · Lower priority

STYLIGHT


All Sources
syslog.host:"WIN-R2T9EJJ3KBT"
Last day
Search
Applied Filters:
json.SourceName : Python ETL
AND
syslog.severity : Error
Field Explorer
17 events
Feb 25 - Feb 26
10
5
0
16:00
18:00
20:00
22:00
26. Feb
02:00
04:00
06:00
08:00
10:00
12:00
14:00
View:
Collapsed Events
Sorting:
Descending
2015-02-26 09:29:00.000 UTC
View Surrounding Events
host : WIN-R2T9EJJ3KBT
log type : syslog
log type : json
tag : tableau_vm
LogglyNotifications:
syslog:
severity: Error
appName: Python_ETL
timestamp: 2015-02-26T09:29:00.000000+00:00
facility: user-level messages
pid: 0
priority: 11
host: WIN-R2T9EJJ3KBT
json:
EventID: 1
ProcessID: 0
EventTime: 2015-02-26 09:29:00
Task: 0
Severity: ERROR
SourceModuleType: im_msvistalog
SourceName: Python ETL
EventType: ERROR
SourceModuleName: eventlog
Hostname: WIN-R2T9EJJ3KBT
Opcode: Info
ThreadID: 0
Keywords: 36028797018963968
SeverityValue: 4
Message:

Thanks for your attention!

STYLIGHT

References

- Google's Cloud Pub/Sub Real-Time Messaging Service Is Now In Public Beta
- Streaming Data Processing with Amazon Kinesis and AWS Lambda
- Google Cloud Dataflow Two Worlds Become a Much Better One

STYLIGHT

Sergii Khomenko
Data Scientist
STYLIGHT GmbH
sergii.khomenko@stylight.com
@lc0d3r

Nymphenburger Straße 86
80636 Munich, Germany

http://daho.am/

STYLIGHT