Data Lake Architectures and Data Science

Diving in

talkdesk DSPT DATA SCIENCE PORTUGAL

Ricardo Teixeira
Senior Data Engineer @ Talkdesk

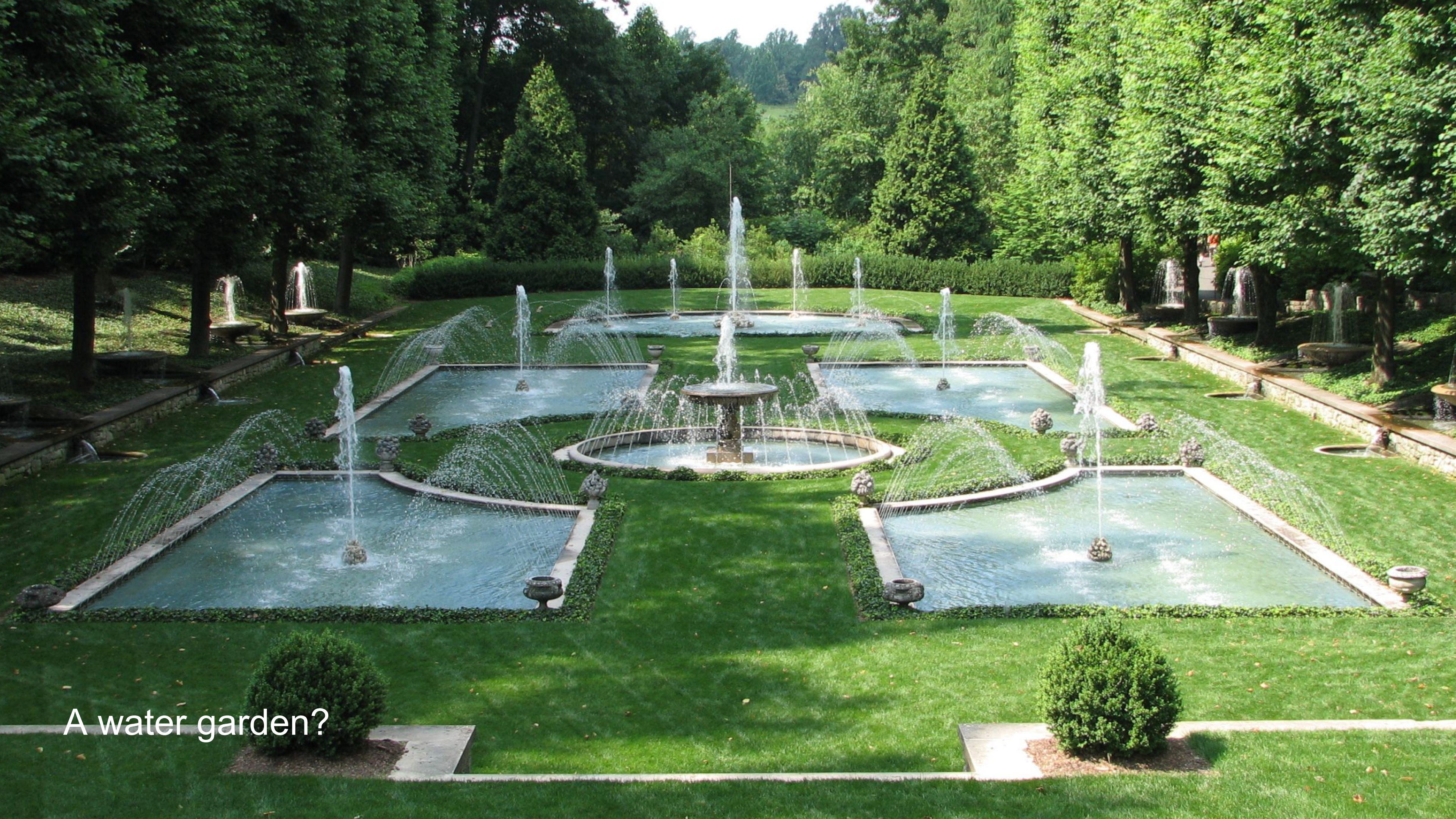
What is a Data Lake?

A data lake is a storage repository that holds an enormous amount of raw or refined data in native format, until it is accessed.

http://lmgtfy.com/?q=data+lake

“If you think of a datamart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples.”

James Dixon, founder and CTO of Pentaho, 2010



A water garden?

Data Lakes ❤ Hadoop

Objective of a Data Lake

Collect everything

A Data Lake contains all data, both raw sources over extended periods of time as well as any processed data.

Dive in anywhere

A Data Lake enables users across multiple business units to refine, explore and enrich data on their terms.

Flexible access

A Data Lake enables multiple data access patterns across a shared infrastructure: batch, interactive, online, search, in-memory and other processing engines.

DATA WAREHOUSE	*vs*	DATA LAKE
structured, processed	***DATA***	structured / semi-structured / unstructured, raw
schema-on-write	***PROCESSING***	schema-on-read
expensive for larger 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.
optimized for known relations	***EXPLORING***	optimized for finding unknown relations

Data Lake Architecture key components

Ingestion

High bandwidth data acquisition of structured, semi-structured and unstructured data

Storage

Highly scalable storage layer that supports unstructured and structured data

Data Management

Catalogue and governance of data inside the data lake

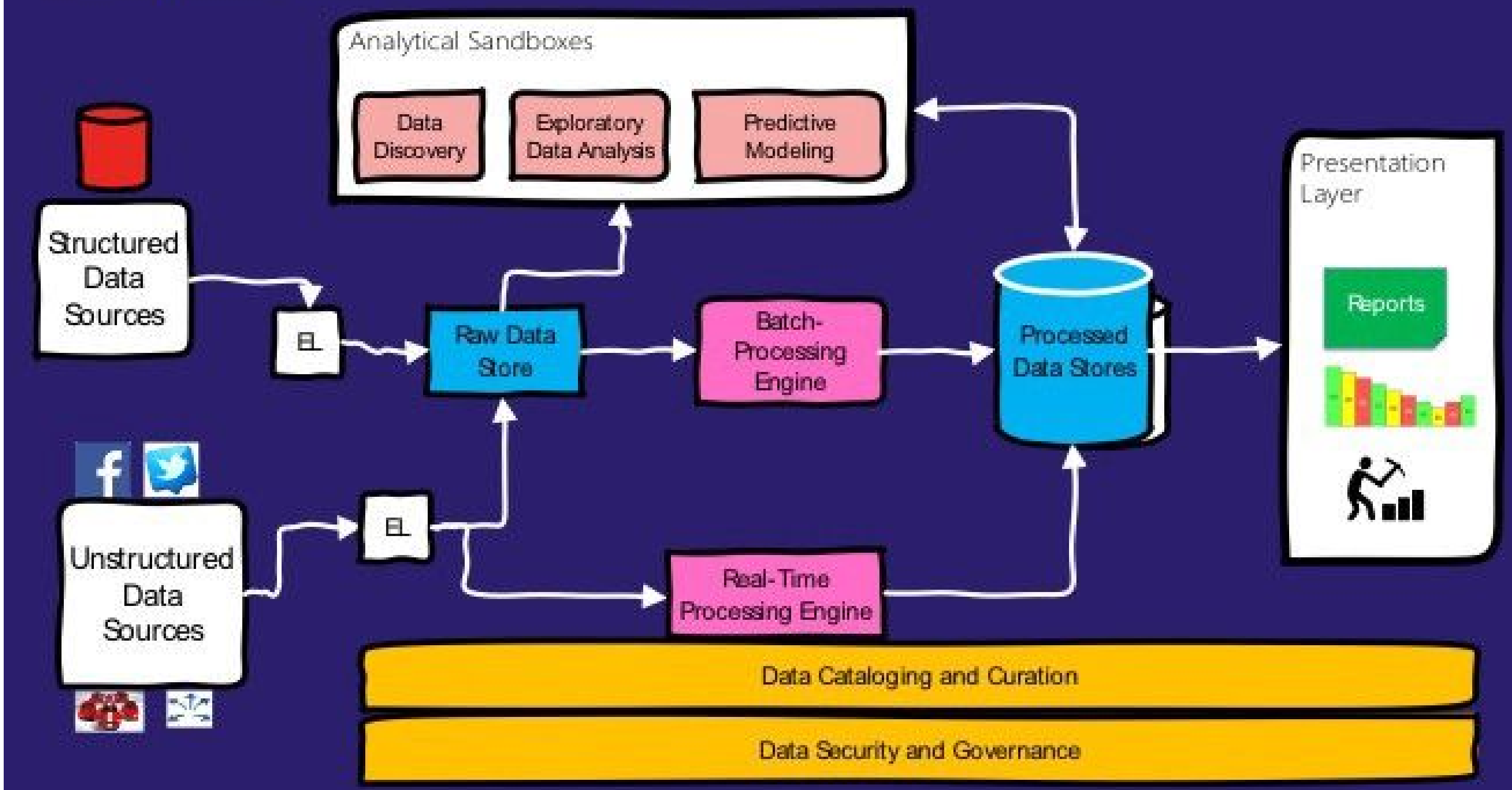
Processing

Batch or real-time distributed highly-scalable processing of data

Discovery and Exploration

Search, query, explore, extract insights

Conceptual Data Lake Architecture



Analytical Sandboxes

Data Discovery

Exploratory Data Analysis

Predictive Modeling

Structured Data Sources

EL

Raw Data Store

Batch-Processing Engine

Processed Data Stores

Presentation Layer

Reports

Unstructured Data Sources

EL

Real-Time Processing Engine

Data Cataloging and Curation

Data Security and Governance

https://medium.com/@prmen/demystifying-data-lake-architecture-30cf4ac8aa07

Data Lake Architecture technologies

Ingestion

- Flume
- Kafka (Connect)
- Apache NiFi
- Amazon Kinesis Firehose
- Logstash
- Sqoop
- (...)

Storage

- HDFS
- S3
- Azure Data Lake Store
- Azure Blog
- Google Cloud Storage

Data Lake Architecture technologies

Management

- Hive Metastore
- Apache Sentry
- Apache Atlas
- Cloudera Navigator

Processing

- Hive
- Spark / Spark Streaming
- Map/Reduce
- YARN
- Google Cloud Dataflow

Exploration

- Presto
- Impala
- Google Big Query
- Amazon Redshift Spectrum & Athena
- Microsoft Data Lake Analytics
- Elasticsearch
- Solr

Pros

Flexibility

Immutable source

Scale

Enable advanced analytics and data science

Facilitates Data Ingestion

Heterogeneous data is welcome

Schema flexibility

Favours Discovery (unknown unknowns)

Cons

Not as mature as Data Warehouses

Harder to query

Loss of trust

Data Swamp

Data governance is harder

- No shared vision of the truth
- Complexity
- Security

Takeaways

Governance and Lineage

Data has to have clear provenance in place and time

Not the end goal

The architecture and systems orbiting around the Data Lake are what makes it powerful

A Data Lake is wherever you store data

Cloud data lakes are the 2018 trend - cloud provider ecosystem is important and direct query is king

Enabler for optimized solutions

Forgo “all in one” tools and use specific, optimized ones that best fit each problem

DS and Data Lakes

Data Science Pipeline

- Data Acquisition and Recording
- Information Extraction and Clean
- Data Integration, Aggregation, and Representation
- Query Processing, Data Modeling, and Analysis
- Interpretation

REAL Data Science Pipeline

- Data Acquisition and Recording
- Information E[illegible] and Clean
- Data Integration, Aggregation, and Representation
- Query Processing, Data Modeling, and Analysis
- Interpretation

Data Science Pipelines using Data Lakes

Pros

- Access to RAW, untainted data
- Throughput
- Timeliness
- Advance analytics
- Ecosystem and tools

Cons

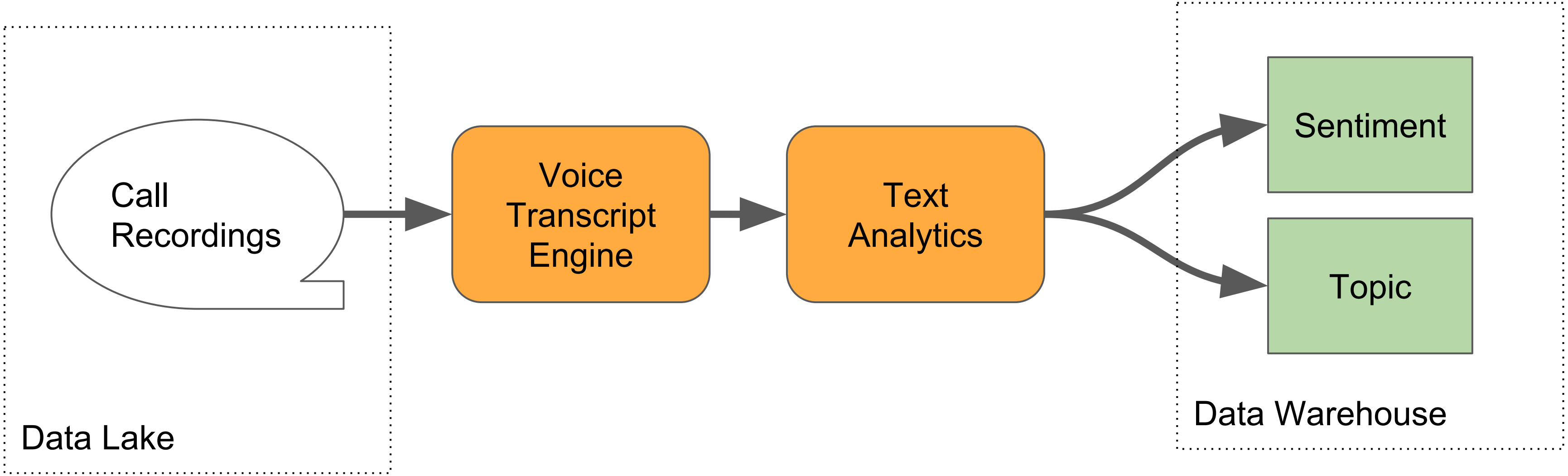
- Heterogeneity and Incompleteness
- Veracity and variety
- Schema less
- Scale
- Privacy concerns
- Less Human Collaboration
- Lack of Curation

Taking advantage of the Data Lake

Scenario 1 - Talkdesk Voice Transcriptions

Scenario 2 - Nasa Image Video Library

Talkdesk Voice Transcriptions



Call Recordings

Data Lake

Voice Transcript Engine

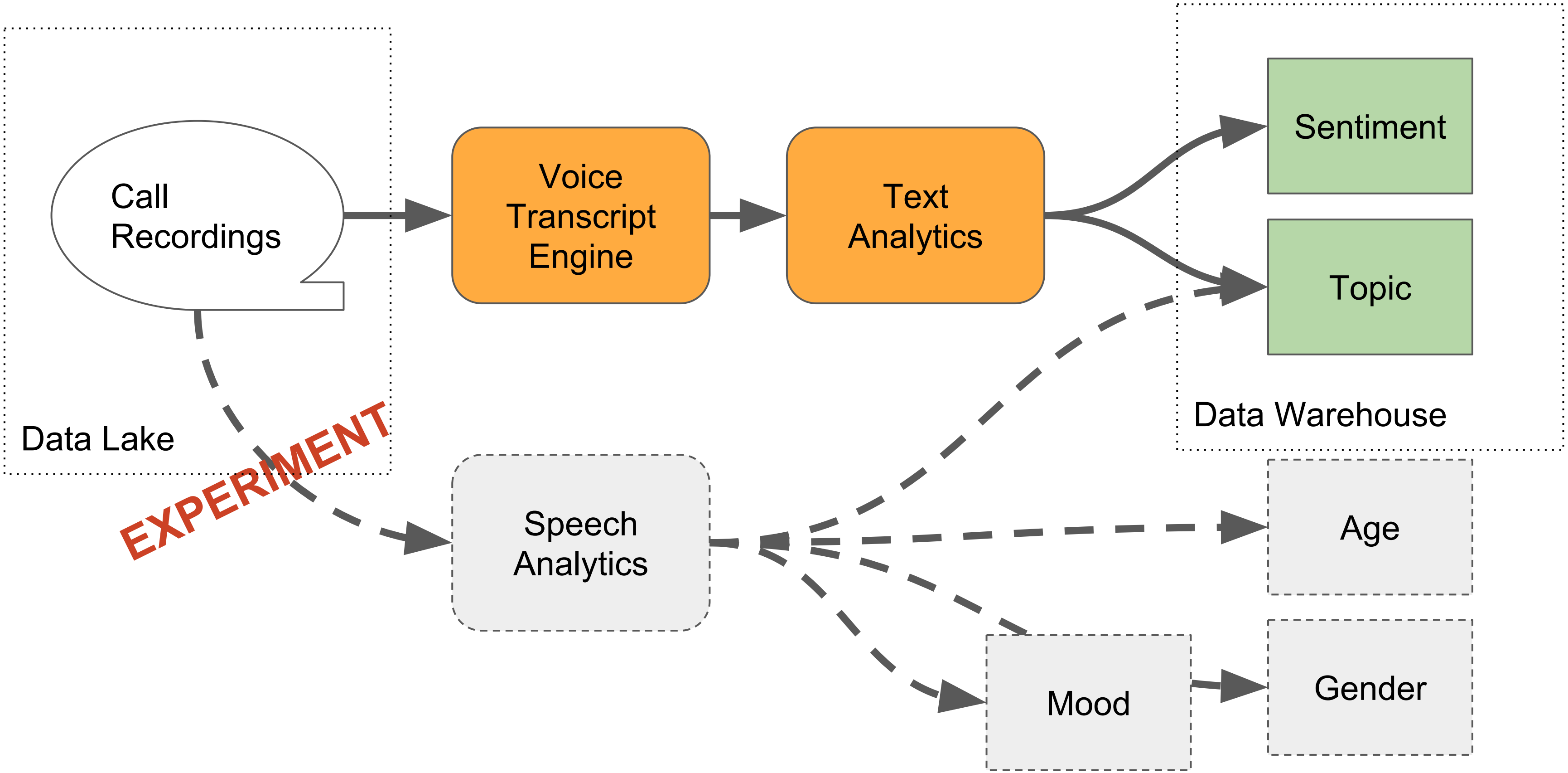
Text Analytics

Sentiment

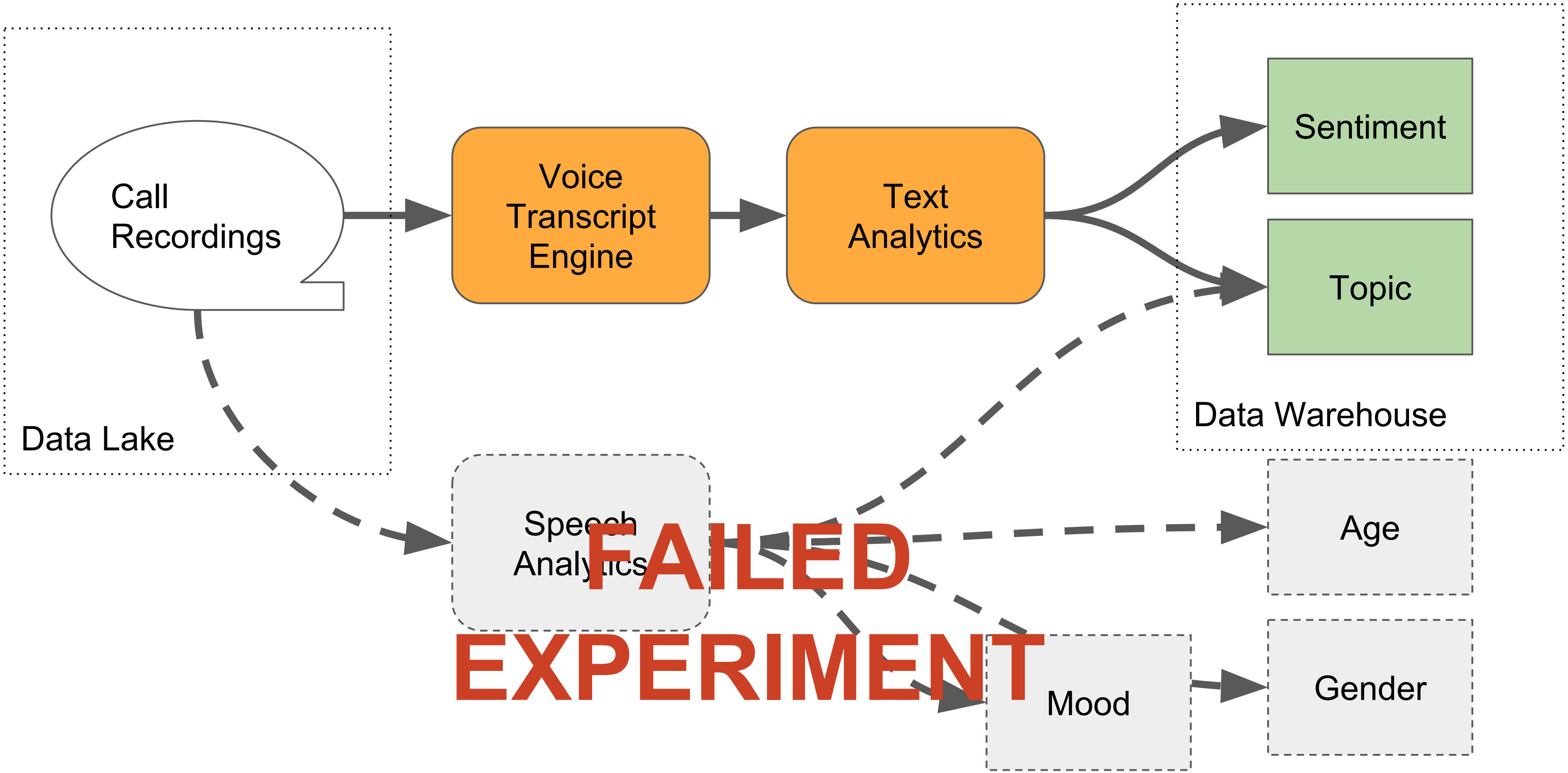
Topic

Data Warehouse

Talkdesk Voice Transcriptions



Talkdesk Voice Transcriptions



Nasa Video Library

NASA Image and Video Library

Search for ... (e.g. "Orion")

☑ Images ☑ Videos ☑ Audio

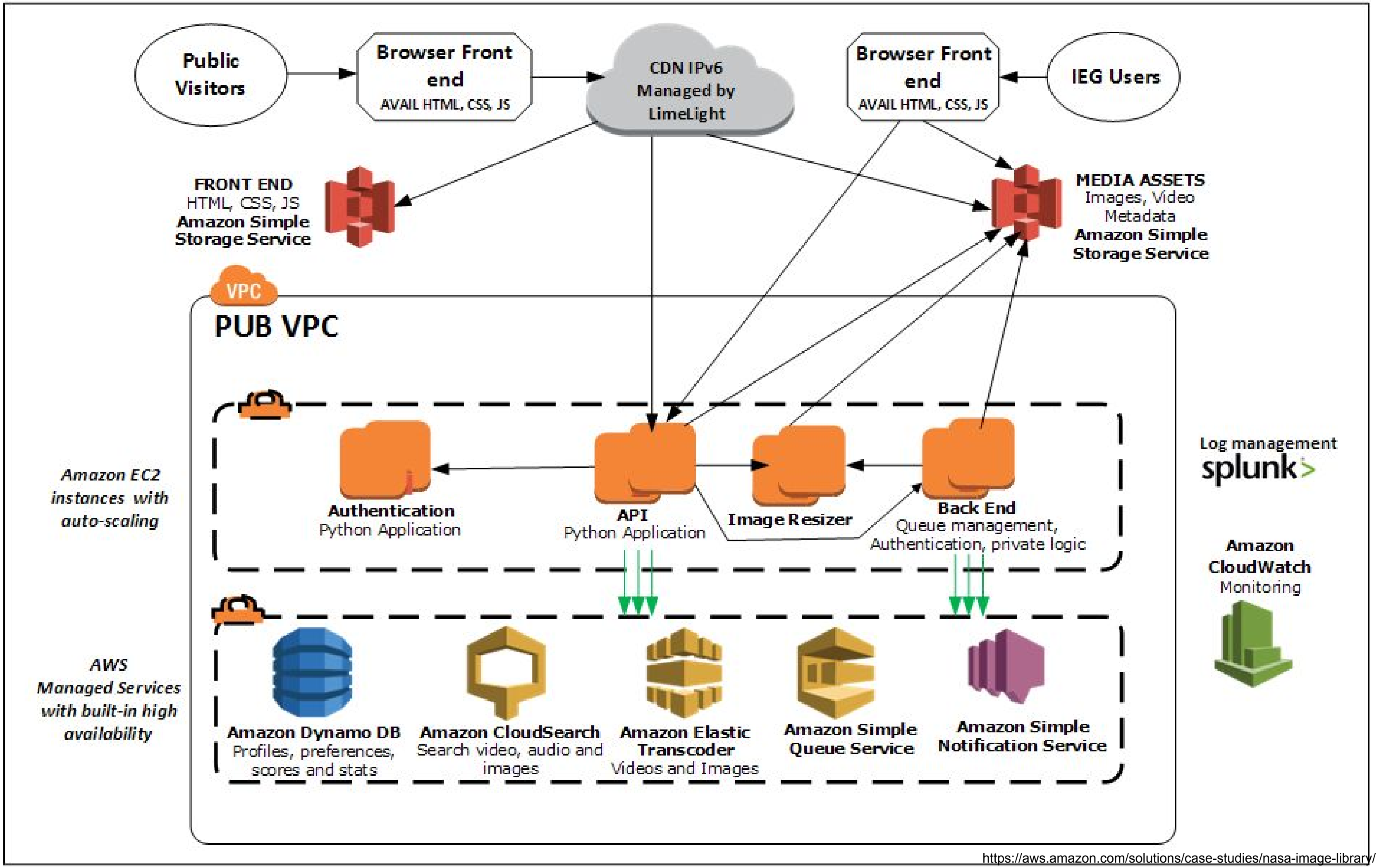
images.nasa.gov

Nasa Video Library

“One-stop shop consisted of essentially “scraping” content from the different (institutional) sites, bringing it together in one place, and layering a search engine on top.”

Nasa Video Library

- Launched in 2017
- Access to images, videos and audio
- In the cloud - AWS
- S3 as the Data Lake
- Access to the metadata associated with each asset
- API for automated uploads of new content



Public Visitors

Browser Front end
AVAIL HTML, CSS, JS

CDN IPv6 Managed by LimeLight

Browser Front end
AVAIL HTML, CSS, JS

IEG Users

FRONT END
HTML, CSS, JS
Amazon Simple Storage Service

MEDIA ASSETS
Images, Video
Metadata
Amazon Simple Storage Service

VPC

PUB VPC

Amazon EC2 instances with auto-scaling

Authentication
Python Application

API
Python Application

Image Resizer

Back End
Queue management, Authentication, private logic

AWS Managed Services with built-in high availability

Amazon Dynamo DB
Profiles, preferences, scores and stats

Amazon CloudSearch
Search video, audio and images

Amazon Elastic Transcoder
Videos and Images

Amazon Simple Queue Service

Amazon Simple Notification Service

Log management
splunk>

Amazon CloudWatch
Monitoring

https://aws.amazon.com/solutions/case-studies/nasa-image-library/

Nasa Video Library

- Easy access
- Scalable
- Built-in Evolution capabilities
- Democratizing access to data

Nasa Video Library

> “We now have an agile, scalable foundation on which to do all kinds of amazing things. Much like with the exploration of space, we’re just starting to imagine all that we can do with it.”

Bryan WallsImagery Experts Deputy Program Manager, NASA

Wrapping it all together

Data Lakes are shiny and cool, but use with caution

Organizations can use use Data Lakes to enable unbounded Data Science exploration, beyond the limits of traditional Data Warehouses

They enable the exploration of massive datasets of RAW, unfiltered data using purpose driven tools

Data Governance is essential for a successful Data Lake implementation

Thank you

Ricardo Teixeira

ricardo.teixeira@talkdesk.com | rteixeira.eu

talkdesk

DSPT
DATA SCIENCE PORTUGAL

"This quote was taken out of context." -- Randall Munroe