

Building a Modern Big Data & Advanced Analytics Pipeline

(Ideas for building UDAP)



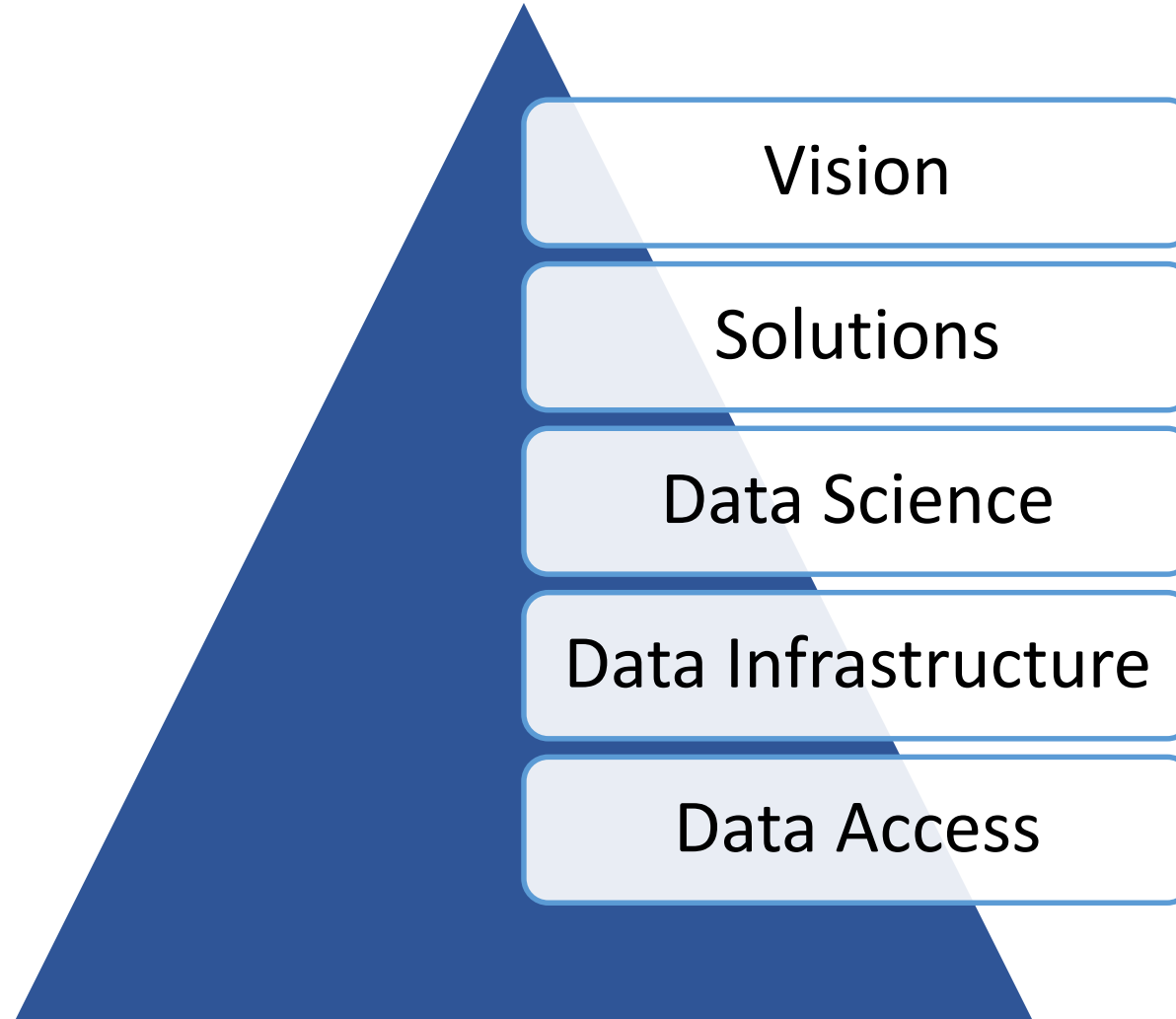
- Emerging technology firm focused on helping enterprises build breakthrough software solutions
- Building software solutions powered by disruptive enterprise software trends
 - Machine learning and data science
 - Cyber-security
 - Enterprise IOT
 - Powered by Cloud and Mobile
- Bringing innovation from startups and academic institutions to the enterprise
- Award winning agencies: Inc 500, American Business Awards, International Business Awards

Agenda

- The principles of big data and advanced analytics pipelines
- Some inspiration
- Capabilities
- Building a big data and advanced analytics pipeline

The principles of an enterprise big data infrastructure

Data Needs



There are only a few
technology choices....

Big Data Landscape 2016

Infrastructure



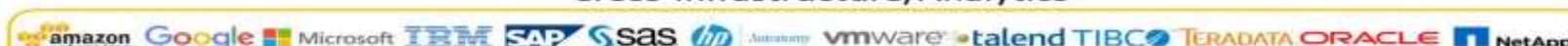
Analytics



Applications



Cross-Infrastructure/Analytics



Open Source



Data Sources & APIs



© Matt Turck (@mattturck), Jim Hao (@jimhao), & FirstMark Capital (@firstmarkcap)

FIRSTMARK

Some inspiration....

Data Access

Data Fetching:
Falcor(<https://github.com/Netflix/falcor>)

Data Streaming: Apache Kafka
(<http://kafka.apache.org/>)

Federated Job Execution
Engine:
Genie(<https://github.com/Netflix/genie>)

Data Infrastructure

Data Lakes: Apache Hadoop
(<http://hadoop.apache.org/>)

Data Compute: Apache Spark
SQL Querying: Presto
(<https://prestodb.io/>)

Data Discovery : Metacat

Data Science

Multidimensional analysis:
Druid (<http://druid.io/>)

Data Visualization: Sting
Machine learning: Scikit-
learn([http://scikit-
learn.org/stable/](http://scikit-learn.org/stable/))

Tools & Solutions

Netflix big data portal

Hadoop Search:
Inviso(<https://github.com/Netflix/inviso>)

Workflow visualization
(<https://github.com/Netflix/Lipstick>)

Netflix Big Data Portal

NETFLIX Big Data Portal

etse

Inbox

Log Out

Home - Query

Dashboard

Schema Search

S3 Browser

Automatic/UC4

Notebooks

Schema Browser

location

filter

prodhive

testhive

s3

segisthus

rds

redshift

teradata

PRESTO Untitled Query

1 select count (*) from dse.ob_exp_allocation_plan_d;

2

Presto

prodhive

RUN and enter or F8

Save

Clear

Show Options

Query History

Data Viewer

Recent Queries

Status Filter: All

Keyword Filter:

allocation_info

12:35

select * from dse.location_d;

12:25

select * from dse.location_rollup_d;

12:27

Bulk Delete Query History...

Page Size: 25

1

Saved Queries

Owned By You

allocation_info

search

Shared With You

Nobody has shared any queries with you.

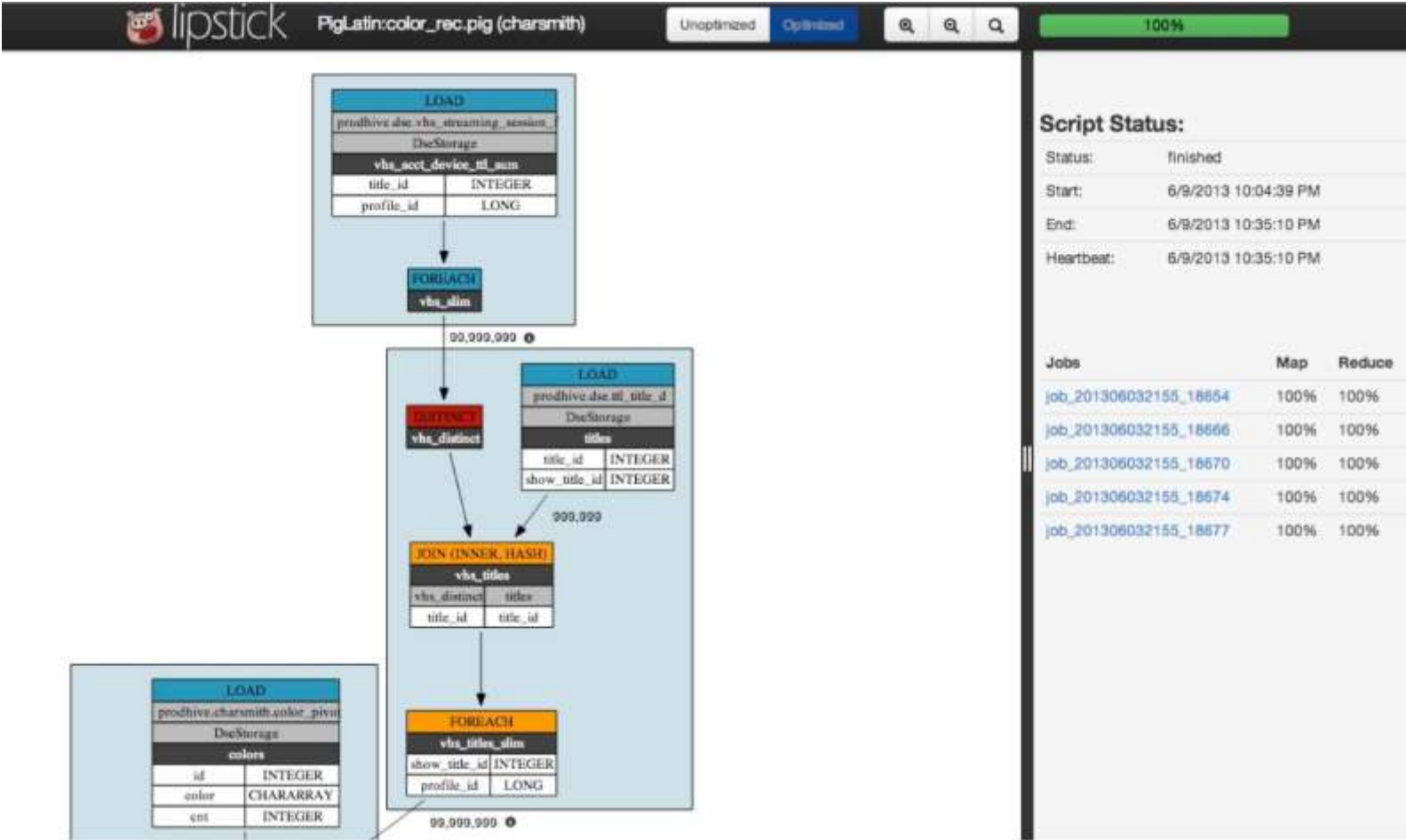
search

Public

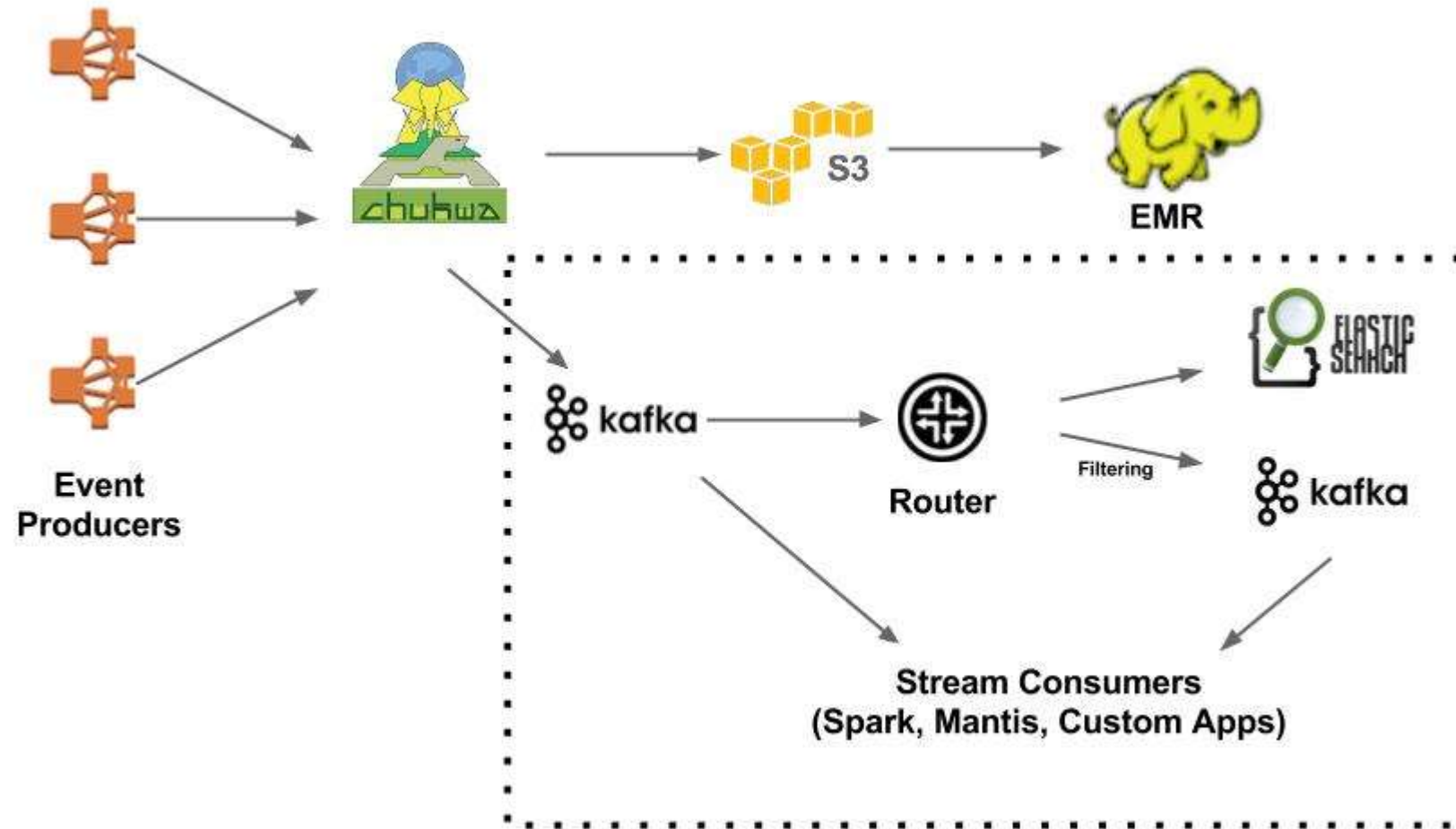
search

Tellago

Netflix Lipstick



Netflix Data Pipeline



Data Access

Data Fetching:

GraphQL(<https://facebook.github.io/react/blog/2015/05/01/graphql-introduction.html>)

Data Streaming: Apache Kafka (<http://kafka.apache.org/>)

Data Infrastructure

Data Lakes: Apache Hadoop (<http://hadoop.apache.org/>)

Data Compute: Apache Spark
SQL Aggregation: Apache
Crunch(<https://crunch.apache.org/>)

Fast Data Access: Apache
Cassandra(<http://cassandra.apache.org/>)

Workflow Manager
:Luigi(<https://github.com/spotify/luigi>)

Data Transformation: Apache
Falcon(<http://hortonworks.com/hadoop/falcon/>)

Data Science


Data Visualization: Sting

Machine learning: Spark
MLib(<http://scikit-learn.org/stable/>)

Data Discovery: Raynor

Tools & Solutions

Hadoop Search:
Inviso(<https://github.com/Netflix/inviso>)

 Spotify | RAYNOR

DATASETS

HELP

Datasets

Shares

Only show certified datasets

AggregatedShares

Rolls up data from SharesCleaned into resource/country/product/platform groups

SharesCleaned

Returns cleaned up data from ClientEventCleaned related to share information

Uses the standardize_source and parse_share_json methods below to standardize disparate types of share logging present in the ClientEvent datasource.

This job is map-only because it is only parsing ClientEvent logs.

Data Access

Data Streaming: Apache Kafka
(<http://kafka.apache.org/>)

Data Fetching:
GraphQL(<https://facebook.github.io/react/blog/2015/05/01/graphql-introduction.html>)

Data Infrastructure

Data Lakes: Apache Hadoop
(<http://hadoop.apache.org/>)

Data Compute: Apache Spark(<http://www.project-voldemort.com/voldemort/>)

Fast Data Access:
Voldemort(<http://cassandra.apache.org/>)

Stream Analytics : Apache Samza(<http://samza.apache.org/>)

Real Time Search : Zoie
(<http://jvasoze.github.io/zoie/>)

Data Science

Multidimensional analysis:
Druid (<http://druid.io/>)

Data Visualization: Sting

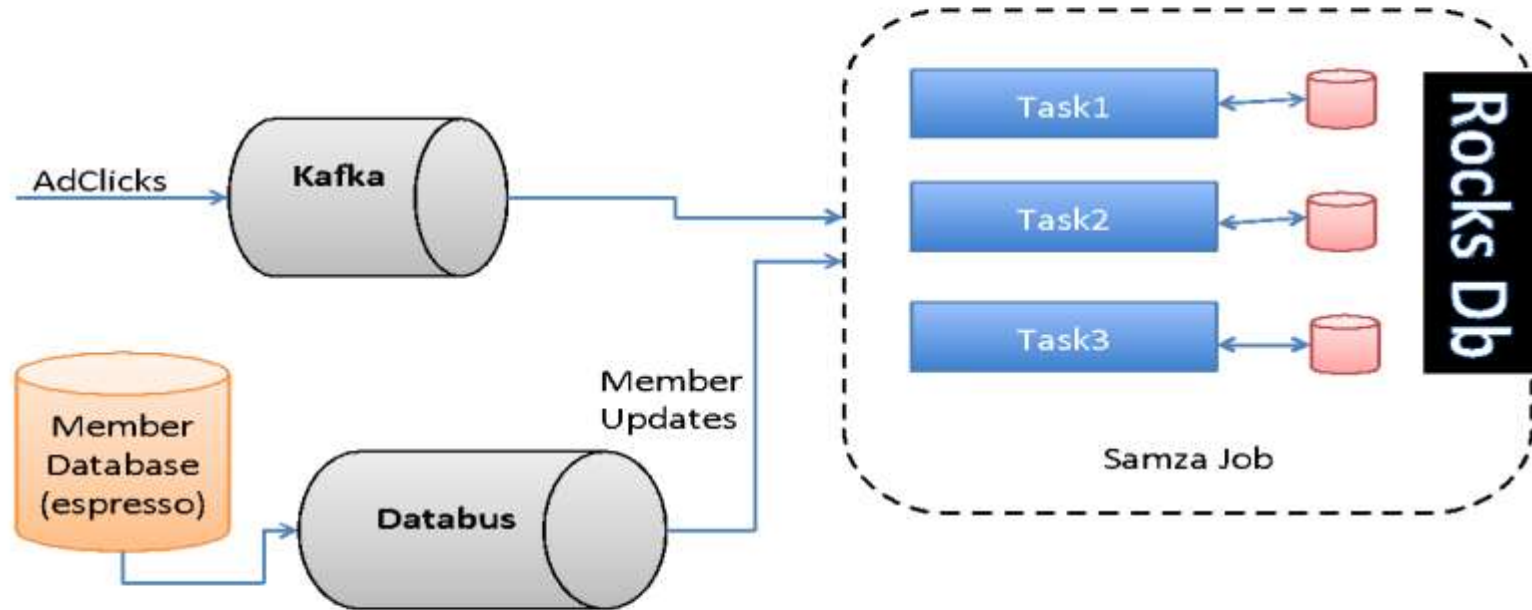
Machine learning: Scikit-learn(<http://scikit-learn.org/stable/>)

Data Discovery: Raynor

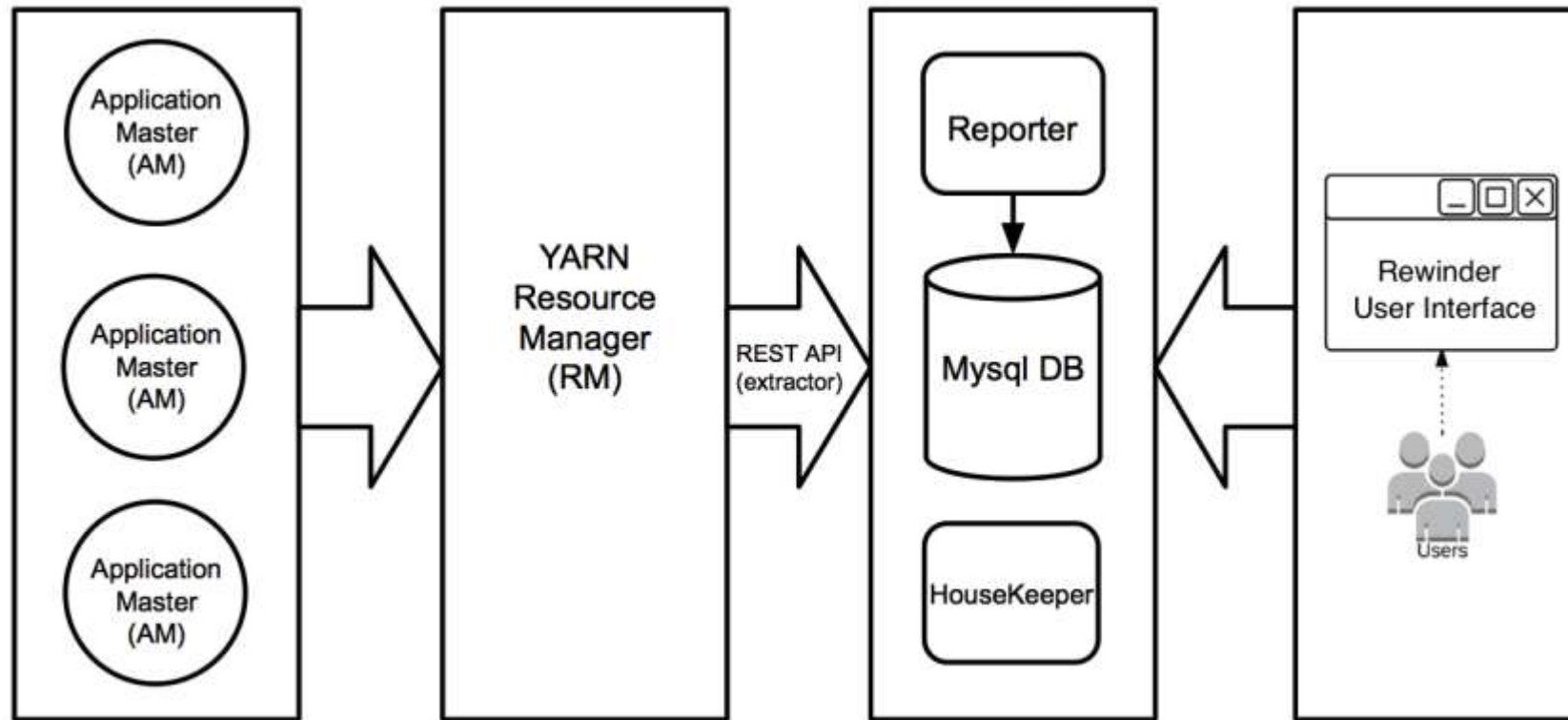
Tools & Solutions

Hadoop Search:
Inviso(<https://github.com/Netflix/inviso>)

LinkedIn Stream Data Processing



LinkedIn Rewinder



Data Access

Data Fetching:

GraphQL(<https://facebook.github.io/react/blog/2015/05/01/graphql-introduction.html>)

Data Streaming: Apache Kafka
(<http://kafka.apache.org/>)

Data Infrastructure

Data Lakes: Apache Hadoop/HBase
(<http://hadoop.apache.org/>)

Data Compute: Apache Spark

Data Transformation: Apache Pig(<http://hortonworks.com/hadoop/falcon/>)

Stream Analytics: Apache Storm
(<http://storm.apache.org/>)

Data Science

Multidimensional analysis: Druid (<http://druid.io/>)

Data Visualization: Sting

Machine learning: Spark MLib(<http://scikit-learn.org/stable/>)

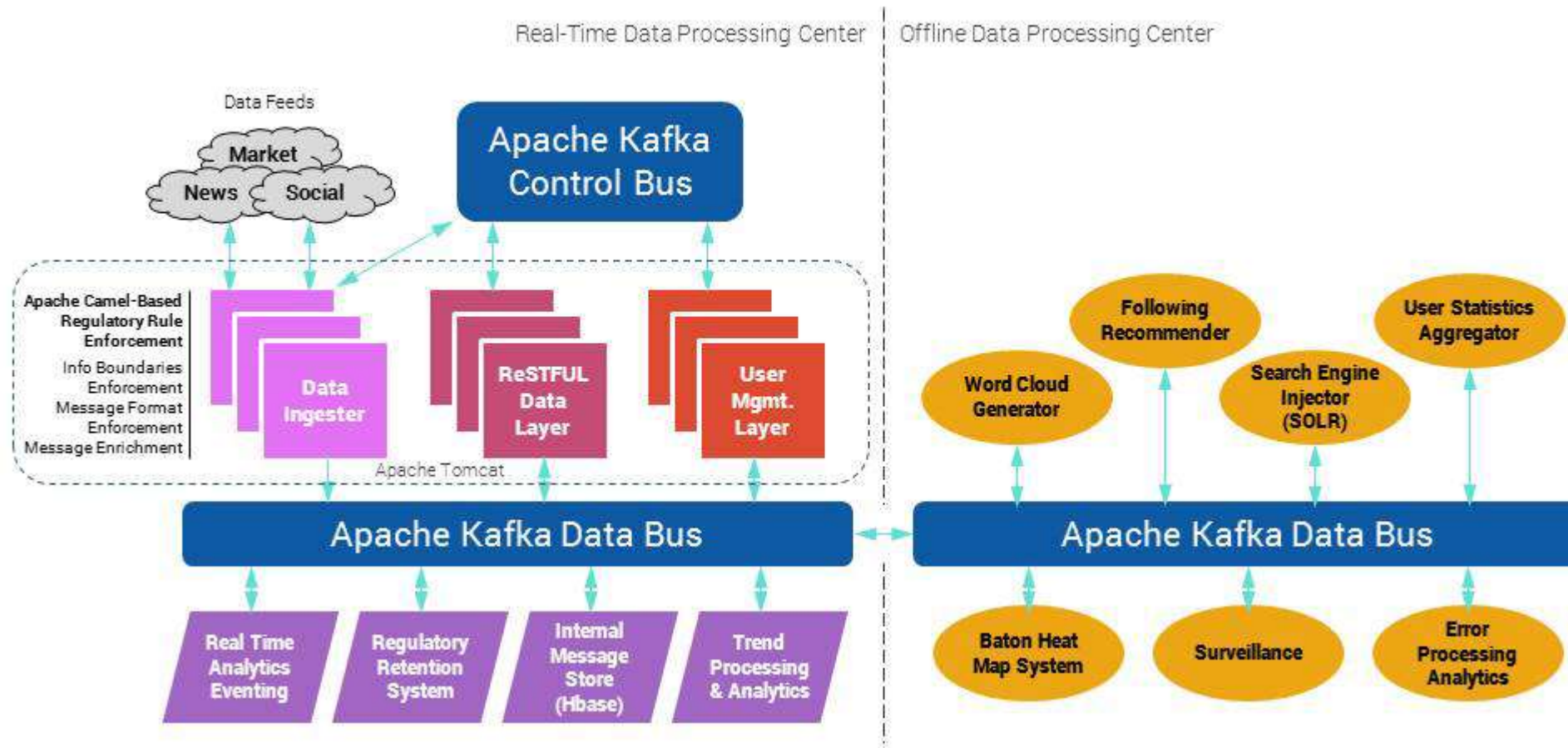
Data Discovery: Custom data catalog

Tools & Solutions

Secure data exchange: Symphony
(<http://www.goldmansachs.com/what-we-do/engineering/see-our-work/inside-symphony.html>)

Goldman Sachs Data Exchange Architecture

Symphony Topology



Capabilities of a big data pipeline

Data Access....

Goals

- Provide the foundation for data collection and data ingestion methods at an enterprise scale
- Support different data collection models in a consistent architecture
- Incorporate and remove data sources without impacting the overall infrastructure

Foundational Capabilities

- On-demand data access
- Batch data access
- Stream data access
- Data transformation

On-Demand Data Access

Best Practices

- Enable standard data access protocols for line of business systems
- Empower client applications with data querying capabilities
- Provide data access infrastructure building blocks such as caching across business data sources

Interesting Technologies

- GraphQL(<https://facebook.github.io/react/blog/2015/05/01/graphql-introduction.html>)
- Odata(<http://odata.org>)
- Falcor(<http://netflix.github.io/falcor/>)

Batch Data Access

Best Practices

- Enable agile ETL models
- Support federated job processing

Interesting Technologies

- Genie(<https://github.com/Netflix/genie>)
- Luigi(<https://github.com/spotify/luigi>)
- Apache Pig(<https://pig.apache.org/>)

Stream Data Access

Best Practices

- Enable streaming data from line of business systems
- Provide the infrastructure to incorporate new data sources such as sensors, web streams etc
- Provide a consistent model for data integration between line of business systems

Interesting Technologies

- Apache Kafka(<http://kafka.apache.org/>)
- RabbitMQ(<https://www.rabbitmq.com/>)
- ZeroMQ(<http://zeromq.org/>)
- Many others....

Data Virtualization

Best Practices

- Enable federated aggregation of disparate data sources
- Focus on small data sources
- Enable standard protocols to access the federated data sources

Interesting Technologies

- Denodo(<http://www.denodo.com/en>)
- JBoss Data Virtualization(<http://www.jboss.org/products/datavirt/overview/>)

Data Infrastructure....

Goals

- Store heterogeneous business data at scale
- Provide consistent models to aggregate and compose data sources from different data sources
- Manage and curate business data sources
- Discover and consume data available in your organization

Foundational Capabilities

- Data lakes
- Data quality
- Data discovery
- Data transformation

Data Lakes

Best Practices

- Focus on complementing and expanding our data warehouse capabilities
- Optimize the data lake to incorporate heterogeneous data sources
- Support multiple data ingestion models
- Consider a hybrid cloud strategy (pilot vs. production)

Interesting Technologies

- Hadoop(<http://hadoop.apache.org/>)
- Hive(<https://hive.apache.org/>)
- Hbase(<https://hbase.apache.org/>)
- Spark(<http://spark.apache.org/>)
- Greenplum(<http://greenplum.org/>)
- Many others....

Data Quality

Best Practices

- Avoid traditional data quality methodologies
- Leverage machine learning to streamline data quality rules
- Leverage modern data quality platforms
- Crowdsourced vs. centralized data quality models

Interesting Technologies

- Trifacta(<http://trifacta.com>)
- Tamr(<http://tamr.com>)
- Alation(<https://alation.com/>)
- Paxata(<http://www.paxata.com/>)

Data Discovery

Best Practices

- Master management solutions don't work with modern data sources
- Promote crowd-sourced vs. centralized data publishing
- Focus on user experience
- Consider build vs. buy options

Interesting Technologies

- Tamr(<http://tamr.com>)
- Custom solutions...
- Spotify Raynor
- Netflix big data portal

Data Transformations

Best Practices

- Enable programmable ETLs
- Support data transformations for both batch and real time data sources
- Agility over robustness

Interesting Technologies

- Apache Pig(<https://pig.apache.org/>)
- Streamsets(<https://streamsets.com/>)
- Apache Spark (<http://spark.apache.org/>)

Data Science....

Goals

- Discover insights of business data sources
- Integrate machine learning capabilities as part of the enterprise data pipeline
- Provide the foundation for predictive analytic capabilities across the enterprise
- Enable programmatic execution of machine learning models

Foundational Capabilities

- Data visualization & self-service BI
- Predictive analytics
- Stream analytics
- Proactive analytics

Data Visualization and Self-Service BI

Best Practices

- Access business data sources from mainstream data visualization tools like Excel , Tableau, QlickView, Datameer, etc.
- Publish data visualizations so that they can be discovered by other information workers
- Embed visualization as part of existing line of business solutions

Interesting Technologies

- Tableau(<http://www.tableau.com/>)
- PowerBI(<https://powerbi.microsoft.com/en-us/>)
- Datameer(<http://www.datameer.com/>)
- QlickView(<http://www.qlik.com/>)
- Visualization libraries
-

Predictive Analytics

Best Practices

- Implement the tools and frameworks to author machine learning models using business data sources
- Expose predictive models via programmable APIs
- Provide the infrastructure to test, train and evaluate machine learning models

Interesting Technologies

- Spark Mlib(<http://spark.apache.org/docs/latest/mllib-guide.html>)
- Scikit-Learn(<http://scikit-learn.org/>)
- Dato(<https://dato.com/>)
- H2o.ai(<http://www.h2o.ai/>)
-

Stream Analytics

Best Practices

- Aggregate data real time from diverse data sources
- Model static queries over dynamic streams of data
- Create simulations and replays of real data streams

Interesting Technologies

- Apache Storm(<http://storm.apache.org/>)
- Spark Streaming (<http://spark.apache.org/streaming/>)
- Apache Samza(<http://samza.apache.org/>)
-

Proactive Analytics

Best Practices

- Automate actions based on the output of predictive models
- Use programmatic models to script proactive analytics business rules
- Continuously test and validate proactive rules

Interesting Technologies

- Spark Mlib(<http://spark.apache.org/docs/latest/mllib-guide.html>)
- Scikit-Learn(<http://scikit-learn.org/>)

Solutions....

Enterprise Data Solutions

- Leverage a consistent data pipeline as part of all solutions
- Empower different teams to contribute to different aspects of the big data pipeline
- Keep track of key metrics about the big data pipeline such as time to deliver solutions, data volume over time, data quality metrics, etc

Some Examples

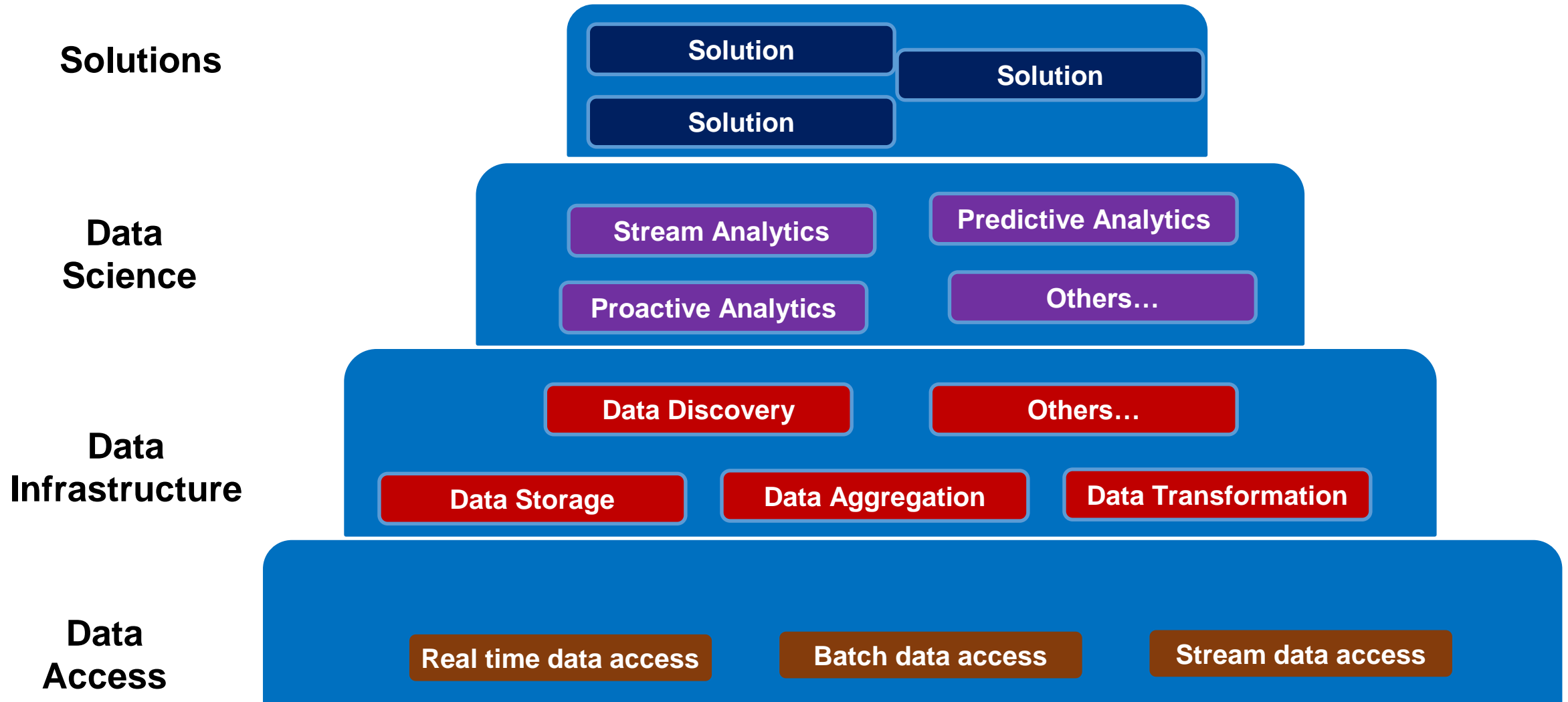
- Data discovery
- Data quality
- Data testing tools
- ...

Other Interesting Capabilities

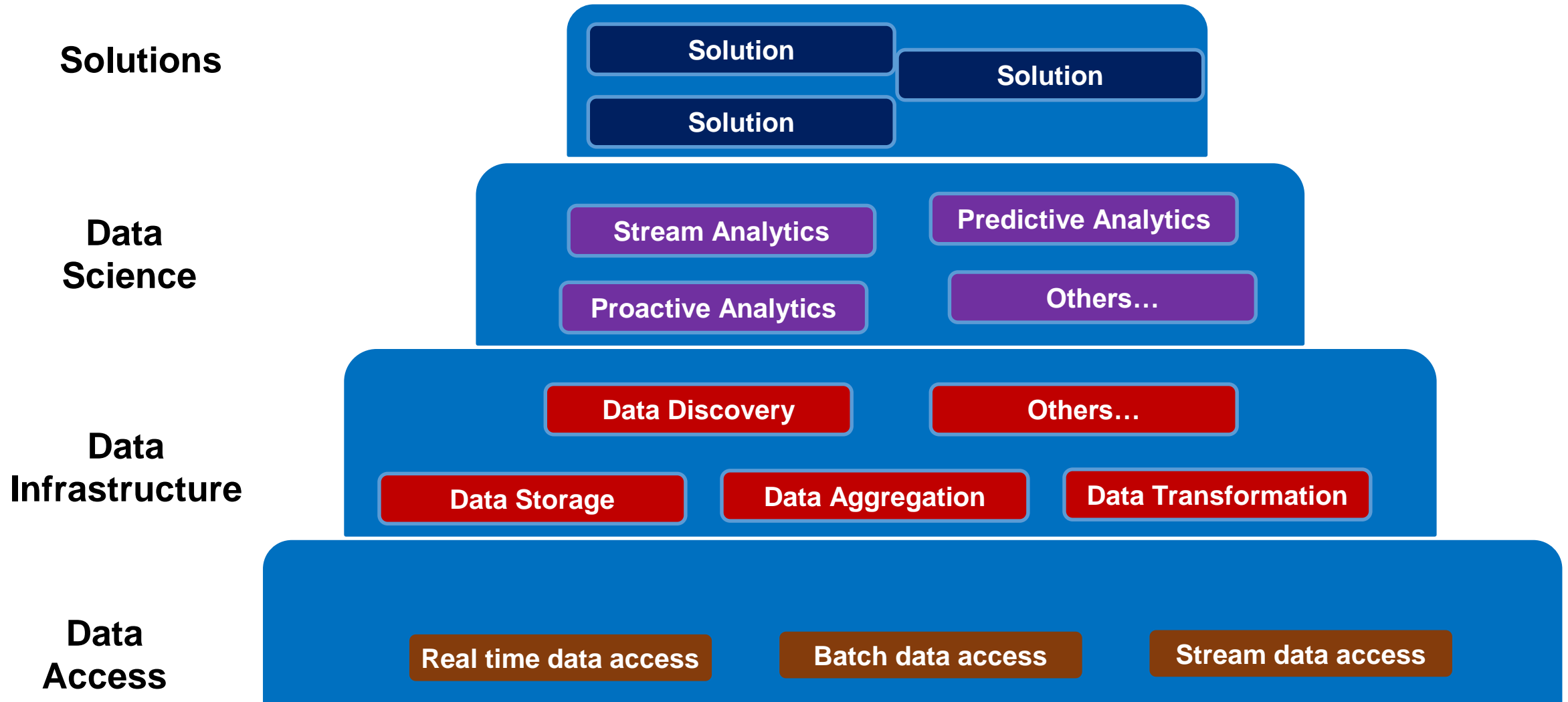
- Mobile analytics
- Embedded analytics capabilities (ex: Salesforce Wave, Workday)
- Aggregation with external sources
- Video & image analytics
- Deep learning
-

Building a big data and advanced analytics pipeline

Infrastructure-Driven



Domain-Driven



Infrastructure-Drives vs. Domain-Driven Approaches

Infrastructure-Driven

- Lead by the architecture team
- Military discipline
- Commitment from business stakeholders

Domain-Driven

- Federated data teams
- Rapid releases
- Pervasive communications

Some General Rules

- Establish a vision across all levels of the data pipeline
- You can't buy everything...Is likely you will build custom data infrastructure building blocks
- Deliver infrastructure and functional capabilities incrementally
- Establish a data innovation group responsible for piloting infrastructure capabilities ahead of production schedules
- Encourage adoption even in early stages
- Iterate

Summary

- Big data and advanced analytics pipelines are based on 4 fundamental elements: data access, data infrastructure, data science, data solutions....
- A lot of inspiration can be learned from the big data solutions built by lead internet vendors
- Establish a common vision and mission
- Start small....iterate....

Thanks

<http://Tellago.com>

Info@Tellago.com

