

Data Analytics Infrastructure

Data Science SG
Nov 2015 Meetup

Le Nguyen The Dat

@lenguyenthebat



Backgrounds

ZALORA Group (2013 – 2014)

- Biggest online fashion retail in South East Asia
- Data Infrastructure & Data Science

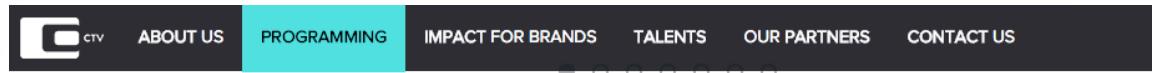
The image is a promotional graphic for ZALORA. It features a grayscale world map focusing on Southeast Asia and Oceania. Overlaid on the map are the names of countries served by ZALORA, arranged in two columns. To the right of the map is the ZALORA logo with the tagline "ASIA'S ONLINE FASHION DESTINATION". Below the logo is a photograph of a man and a woman standing side-by-side. The man is wearing a dark blazer, a green t-shirt, and tan shorts, and is wearing a brown fedora hat. The woman has long, curly hair and is wearing a light blue top and a floral skirt.

MALAYSIA	SINGAPORE
BRUNEI	HONG KONG
TAIWAN	PHILIPPINES
VIETNAM	THAILAND
INDONESIA	AUSTRALIA
NEW ZEALAND	

Backgrounds

Commercialize.TV (2015 –)

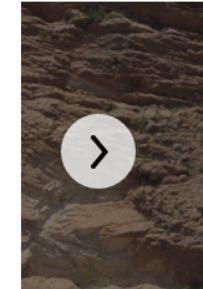
- Multi Channel media network – focusing on China audiences
- Data Infrastructure & Insights



KIDS



FASHION



ACT

your little ones – and you! On our kids network, if you need to keep your kids entertained and reviews, new animation blockbusters and fun, do-it-cts – these are just some of the engaging content t will keep your child's curious mind captivated and inspired.

The ultimate curation of professional fashion show content from our leading industry partners alongside news, gossip, celebrity and hand-chosen emerging creator content. From the runways of New York to Paris, Milan and Shanghai we have exclusive behind the scenes access. We have rights to exclusive shows and simultaneously foster up-and-coming online fashion & beauty talent. The perfect mix of aspirational and accessible content from the East and the West.

Thrilling adrenaline-pui achievements in the grea look-away global mix of athletes alongside first-hi breathe their passion. Sno sublime emersion in the d breadth c

► LUXURY ► FASHION ► ACTION & ADVENTURE
► TRAVEL ► FOOD ► ENTERTAINMENT ► KIDS

Challenges

No central data source:

- Data stored in multiple locations
- Unclear ownership

Data definition and quality:

- Little to none documentation
- Different formula, rules owned by different departments
- Always dirty no matter what

Reporting – Descriptive analytics:

- Immediate needs, automations
- Important to do it right (and quick!)



Data Warehouse

Database Technologies

SQL – Relational Databases:

- MySQL, PostgreSQL
- MS SQL Server, Oracle SQL

NoSQL:

- Redis
- Cassandra
- MongoDB
- DynamoDB (AWS)
- RethinkDB

Map Reduce ecosystem:

- Hadoop: HDFS – Pig – Hive – Hbase
- Spark: RDD – Spork – Shark (Spark SQL) – Hbase-Spark

Massively Parallel Processing (MPP):

- Vertica (HP) - Greenplum (EMC) – Netezza (IBM)
- ParAccel (Amazon Redshift)



Database Technologies

SQL – Relational Databases: ✓

- MySQL, PostgreSQL
- MS SQL Server, Oracle SQL

NoSQL: (?)

- Redis ✗
- Cassandra ✓
- MongoDB ✗
- DynamoDB (AWS) ✗
- Neo4j ✗

Map Reduce ecosystem: ✓

- Hadoop: HDFS – Pig – Hive – Hbase
- Spark: RDD – Spork – Shark (Spark SQL) – Hbase-Spark

Massively Parallel Processing (MPP): ✓

- Vertica (HP) - Greenplum (EMC) – Netezza (IBM)
- ParAccel (Amazon Redshift)

Data Warehouse

Amazon Redshift

aws.amazon.com/redshift

- Cloud-based, Fully managed
- SQL (PostgreSQL 8.0.2)
- On-demand (\$2000/year)
- Scalable (Petabyte-scale)
- FAST! amplab.cs.berkeley.edu/benchmark/

All in ONE place!

- Product information
- Customer information
- Tracking data
- External data sources
(Social Media, 3rd Party datasets)

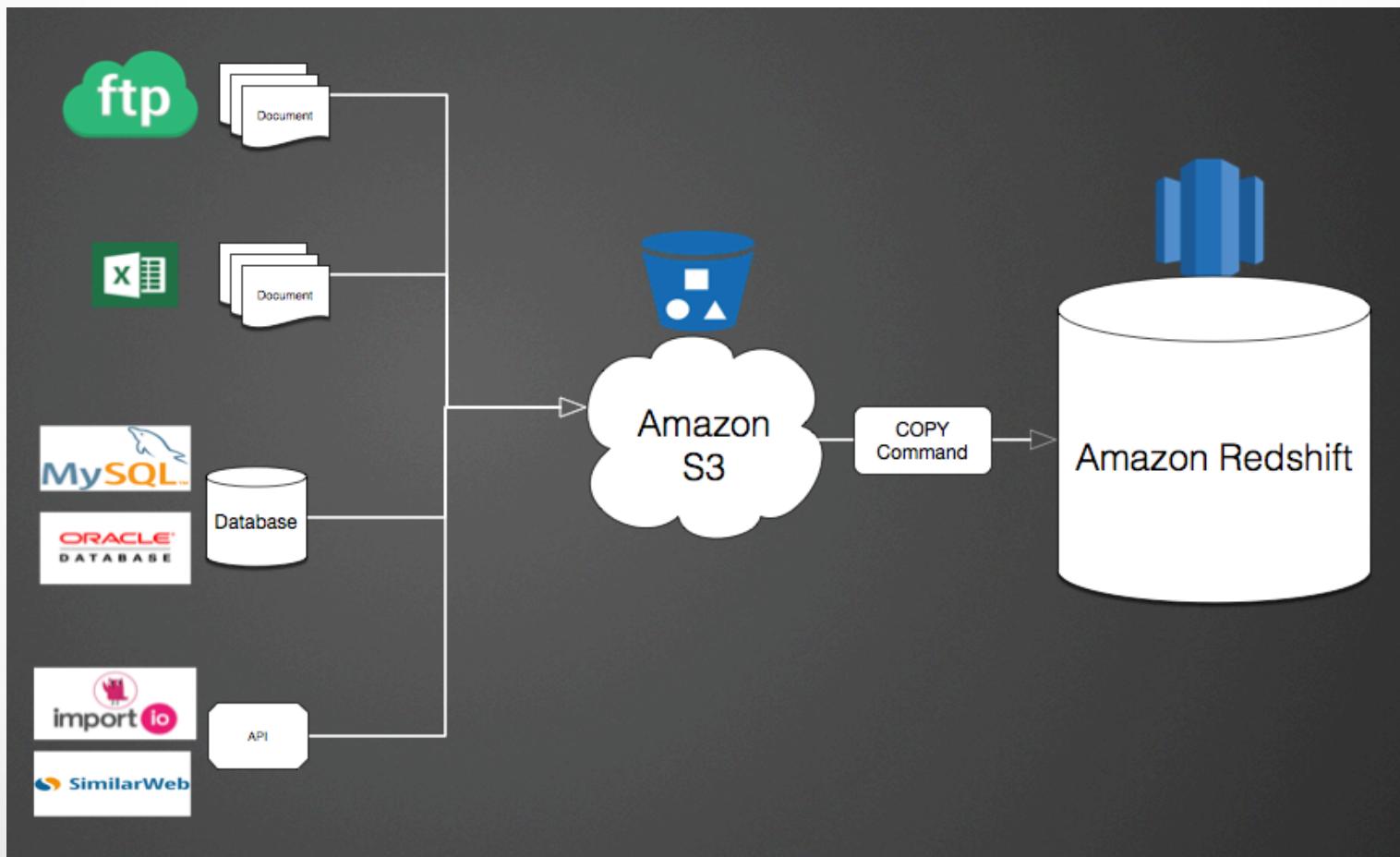


Extract-Transform-Load (ETL)

ETL

Amazon Redshift's COPY command.

docs.aws.amazon.com/redshift/latest/dg/r_COPY.html



ETL

Custom made:

- Simple bash script, python, SQL
- Use cases:
 - Scrappers
 - Excel / CSV imports

Data Pipeline Frameworks:

- Large scale, more complicated
- Examples:
 - Spotify's Luigi – github.com/spotify/luigi
 - Yelp's Mycroft – github.com/Yelp/mycroft

3rd Party Services:

- aws.amazon.com/redshift/partners
 - Flydata
 - Rjmetrics

Applications

Self-services Dashboards

Intermediate users

- SQL / Excel / WYSIWYG Tools
- Re:dash – www.redash.io
 - Open source: github.com/getredash/redash
 - Try it out: demo.redash.io
 - Self-manage & deployment:
 - Docker
 - Pre-baked AMI (Amazon Web Services)
 - Google Cloud Images
 - Supports lots of database types (Redshift, MySQL, PostgreSQL, Big Query, MongoDB...)
 - Users need to know SQL
 - Web-based, collaborative work type

Demo: re:dash data sources usage

<http://demo.redash.io/queries/756>

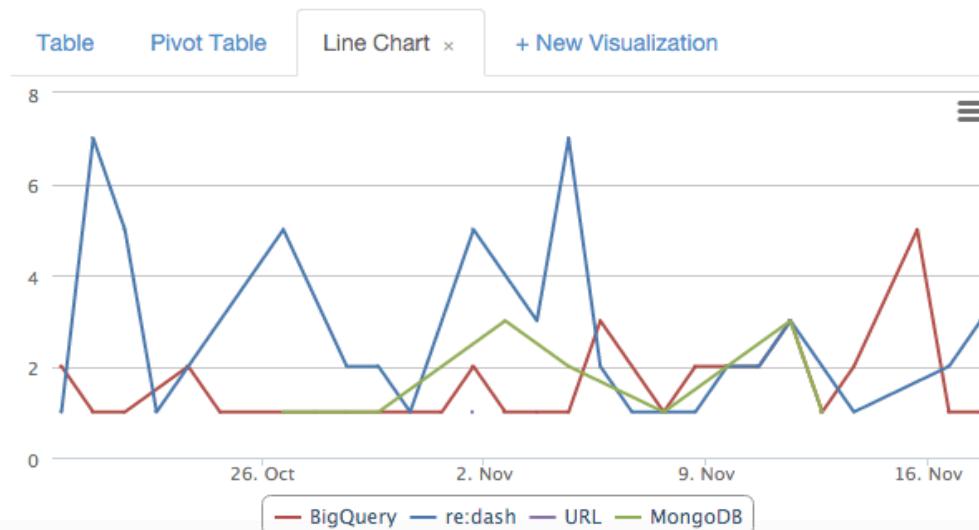
Execute Format SQL Fork Save Search schema...

```
1 SELECT data_sources.name, date(queries.created_at), count(*)
2 FROM queries
3 LEFT JOIN data_sources ON data_sources.id = queries.data_source_id
4 WHERE date(queries.created_at) > CURRENT_DATE - INTERVAL '1 months'
5 GROUP BY data_sources.name, date(queries.created_at)
```

cohort2 cohort_example dashboards data_sources events queries query_results visualizations widgets

Created By Le Nguyen The Dat
Last update 11 minutes ago
Runtime 0s
Rows 52
Refresh Schedule Every 24h
Data Source re:dash

Download Dataset



Demo: NYC Taxis Tip Amounts

<http://demo.redash.io/queries/753>

▶ Execute Format SQL

Fork Save

```
1 SELECT INTEGER(ROUND(FLOAT(tip_amount) / FLOAT(fare_amount) * 100)) tip_pct,
2       count(*) trips
3 FROM [833682135931:nyctaxi.trip_fare]
4 WHERE payment_type='CRD'
5   AND float(fare_amount) > 0.00
6 GROUP BY 1
7 ORDER BY 2 DESC LIMIT 20
```

Created By Le Nguyen The Dat

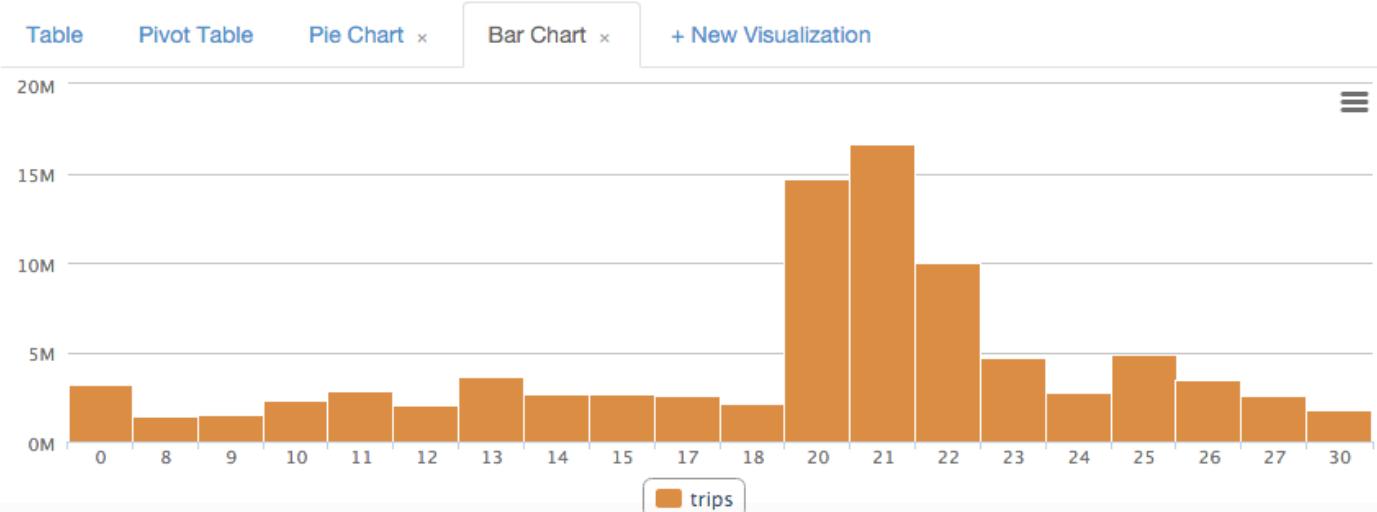
Last update 2 days ago

Runtime 3s

Rows 20

Refresh Schedule Never

Data Source BigQuery



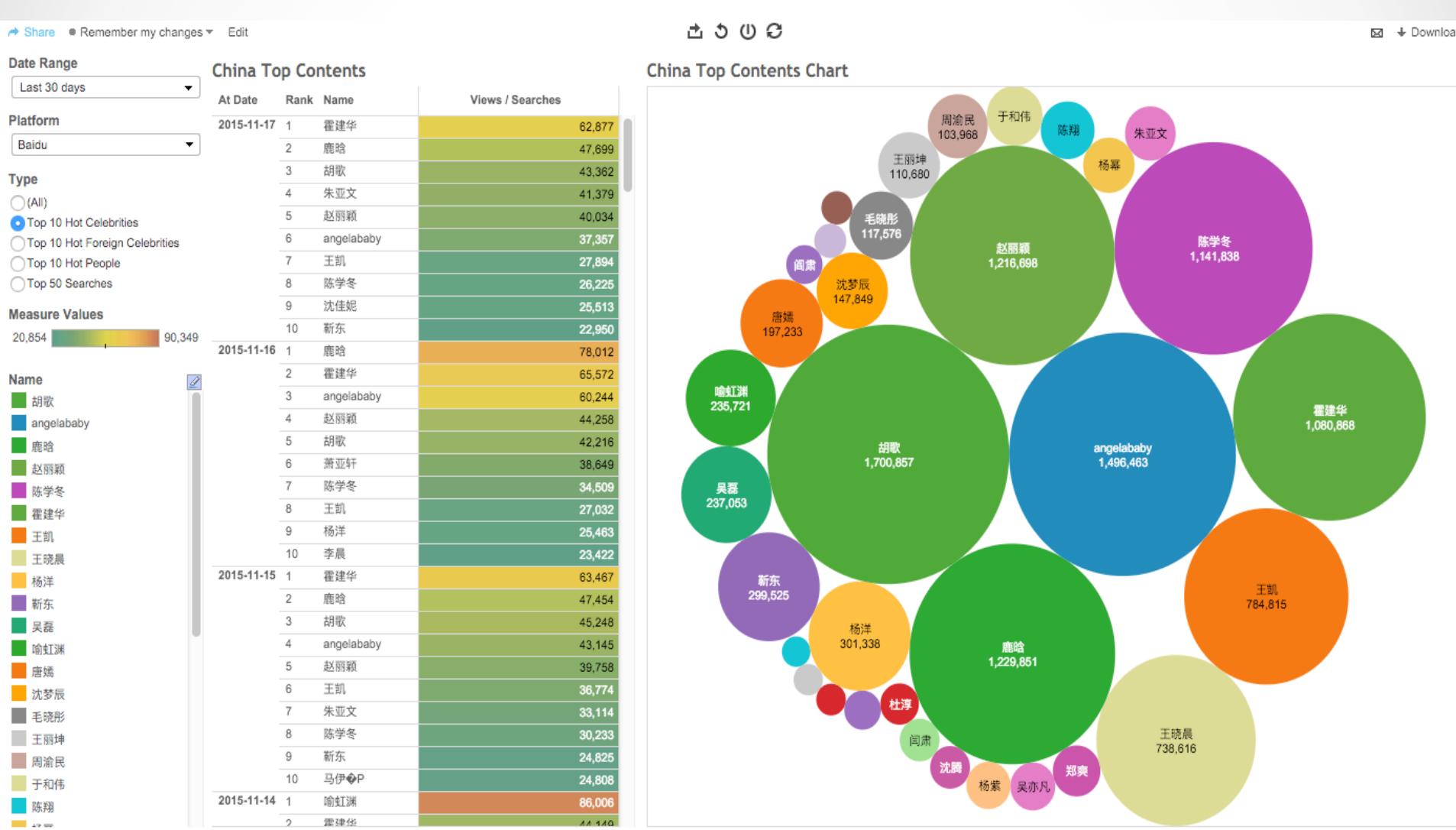
Self-services Dashboards

Intermediate users

- SQL / Excel / WYSIWYG Tools
- Tableau – www.tableau.com
 - Licensed software (14 days trial)
 - Tableau Public (Free: public.tableau.com)
 - Self-host or Tableau host (fully managed)
 - Supports a lot more database types
 - Group, User management – customized access right
 - Drag & Drop software as well as web-based

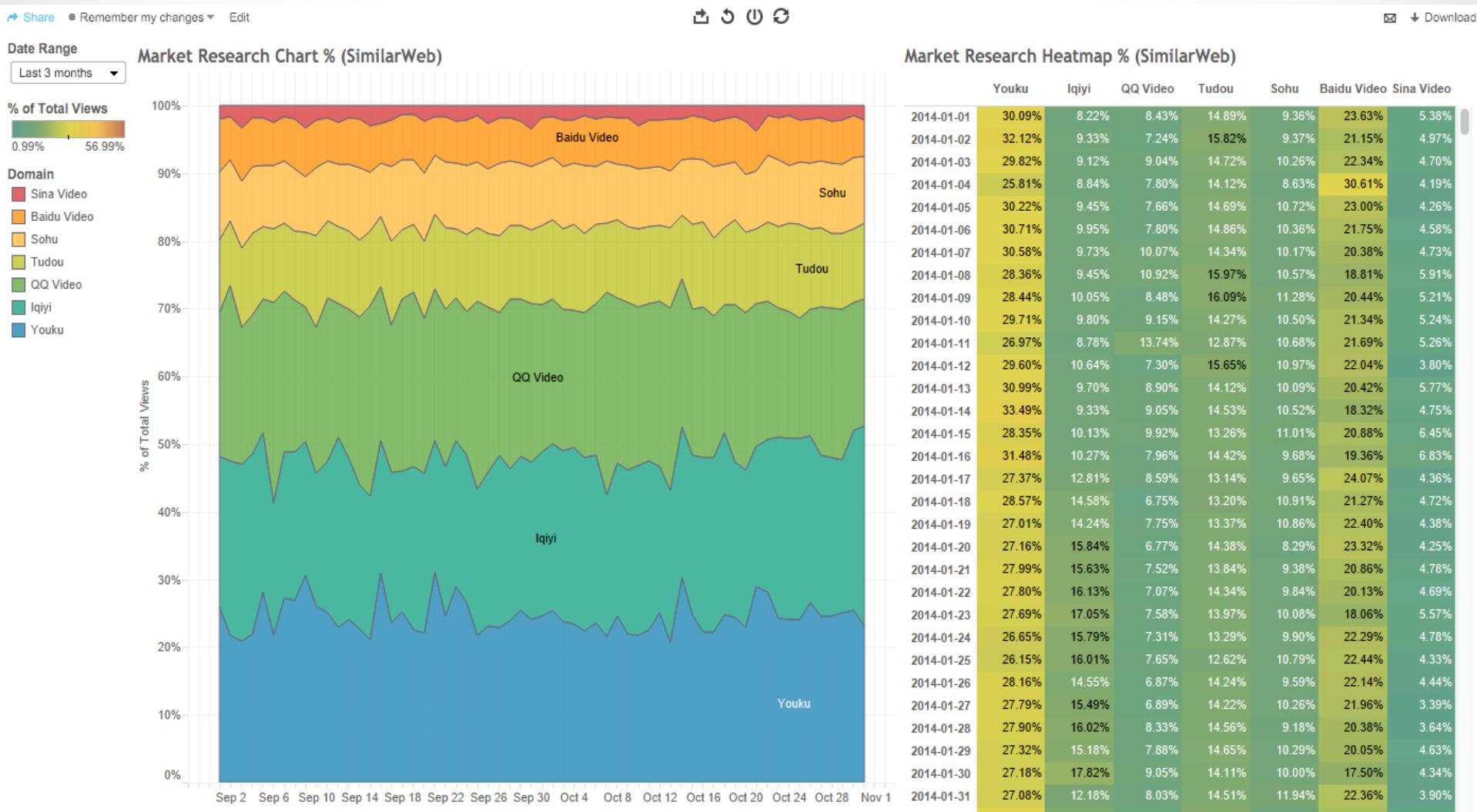
Demo: Social media dashboard

Baidu → Import.io API → Data Warehouse → Tableau



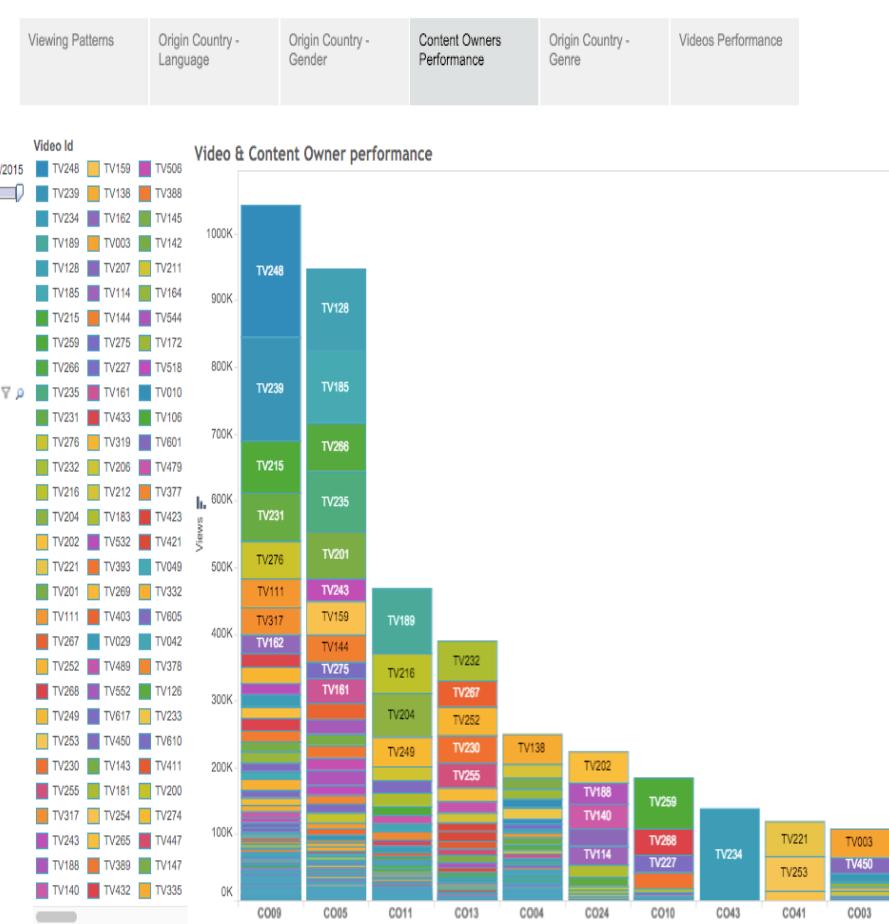
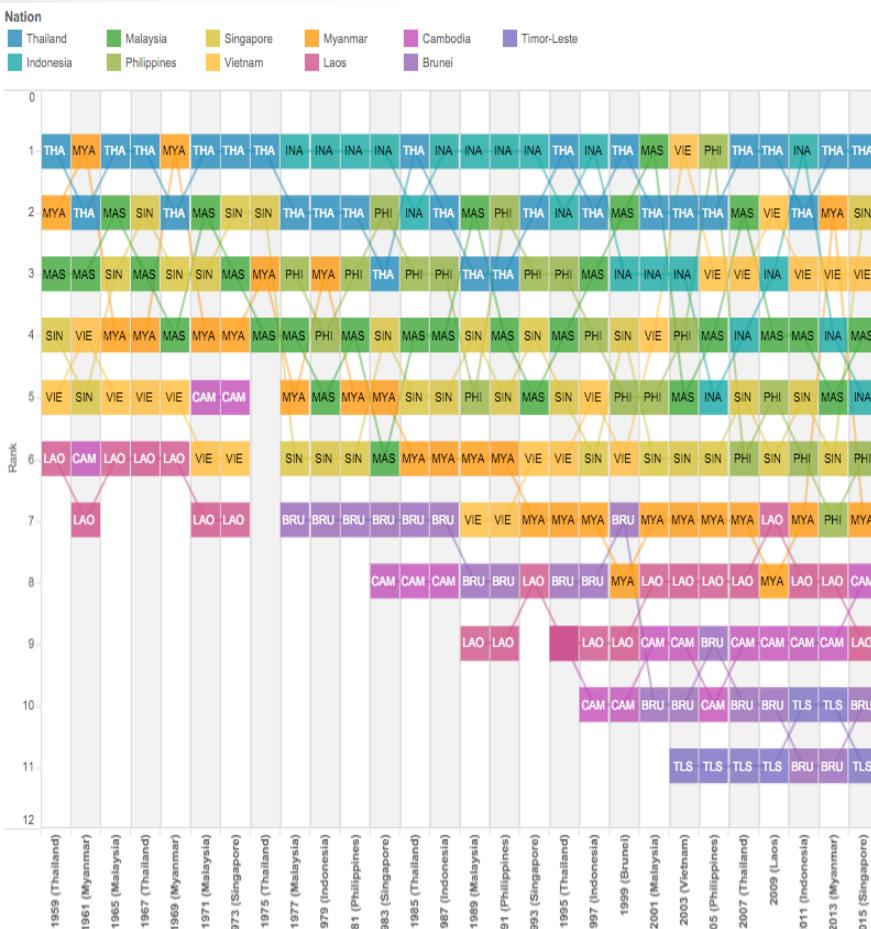
Demo: Market Research - video platform performance

SimilarWeb API → Data Warehouse → Tableau



Others (Tableau Public):

- SEA Games Result history tiny.cc/seagames
 - Rakuten – Viki data challenge tiny.cc/viki-viz



Advanced applications

Advanced users

- Data Warehouse connection (JDBC - PostgreSQL)
- Automated, highly customized reports.
- Data Science:
 - Recommendation engine
 - Predictive modeling
 - Classifications

Advanced applications

Internal reporting tool

Data Warehouse → SQL, Python (Django), JS → Product-Finder

[Black dress | SKU or ID | Tiffany | Atmosphere]

Product Finder

find

SKU SKU Supplier Config Brand Name Product Name

Country Supplier Type Buying Brand Department Season Buying Planning Category Sub Category

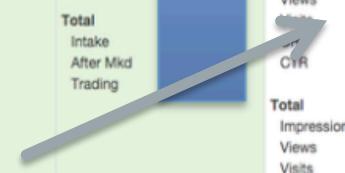
Results per page 10 20 50 100

← 1 2 3 4 5 6 7 8 9 10 ... →

Country		SKU Info	Stock (OMS)	Stock (BOB) Warning!	# Net Sale	Rate of Sale	# Revenue 1	# Margin 1	Webtrekk Data Warning!
AU	 [ONLINE]	<p>SKU ATO48AA66DKN SKU Supplier Config AWS13KD261 ID Catalog Config 125333 Product Name Tiffany Bodycon Dress Color Family black Brand AtmosHere Brand Type Private Label Sub Category Dresses Buying Planning Category Wapp Buying Brand Department N/A Season Sum14Main Supplier Type Main Store</p> <p>AUD \$79.95 NOW AUD \$51.98</p>			<p>414 Last 7 Days Sold Returned/Rejected Cancelled</p> <p>Last 30 Days Sold Returned/Rejected Cancelled</p> <p>Total Sold Returned/Rejected Cancelled</p>	<p>More info</p> <p>Per Day</p> <p>Per Week</p> <p>Per Month</p> <p>(Days Online)</p>	<p>Last 7 Days Intake After Mkd Trading</p> <p>Last 30 Days Intake After Mkd Trading</p> <p>Total Intake After Mkd Trading</p>	<p>Last 7 Days Intake After Mkd Trading</p> <p>Last 30 Days Intake After Mkd Trading</p> <p>Total Intake After Mkd Trading</p>	<p>Last 7 Days Impressions Views Visits CR CTR</p> <p>Last 30 Days Impressions Views Visits CR CTR</p> <p>Total Impressions Views Visits CR CTR</p>

Product Info 

Sales Info 

Tracking Info 

Advanced applications

Recommendation Engine

Data Warehouse → SQL, Python, Haskell → ZALORA Website

The image displays four screenshots of the ZALORA website illustrating a recommendation engine. Each screenshot shows a main product image on the left and a 'Similar products:' section on the right.

- Top Left:** A woman's sleeveless top. The 'Similar products:' section shows four similar tops. Red arrows point from the main image to the recommendation section.
- Top Right:** A black and white strappy sandal. The 'Similar products:' section shows five similar sandals. Red arrows point from the main image to the recommendation section.
- Bottom Left:** A man's blue and white checkered shirt. The 'Similar products:' section shows four similar shirts. Red arrows point from the main image to the recommendation section.
- Bottom Right:** A man's tan chino shorts. The 'Similar products:' section shows four similar pairs of shorts. Red arrows point from the main image to the recommendation section.

In the bottom right screenshot, there is a promotional banner for 'ENDS JUNE 10' and 'EXTRA 20% OFF CODE: MNG20'.

Conclusions

Team & Technology stack

- Small team of 1-4 programmers



- Amazon Web Services

- *No upfront cost*
 - *Low maintenance*
 - *Scalability*
 - *Integrations*



- Shell Scripts, Python, Haskell, D3.js



- Unix, Open-source technologies



Takeaways

- (Good) data infrastructure is important:
 - Build it first (before you hire a data scientist!)
 - Build it right: stable – fast – scalable.
- There is no silver bullet:
 - Understand what you need
 - Always do more research
- Data infrastructure is NOT that hard!
 - Utilize existing, modern technologies
 - Avoid old, proprietary technology that were built for the 90s!

References

Engineering Blogs and Tutorials

- <https://blog.asana.com/2014/11/stable-accessible-data-infrastructure-startup/>
- https://medium.com/@samson_hu/building-analytics-at-500px-92e9a7005c83
- <https://engineering.pinterest.com/blog/powering-interactive-data-analysis-redshift>
- <http://engineering.ifttt.com/data/2015/10/14/data-infrastructure/>
- <https://blog.rjmetrics.com/2015/10/15/the-data-infrastructure-meta-analysis-how-top-engineering-organizations-built-their-big-data-stacks/>
- <https://www.youtube.com/watch?v=reQtXquDpzo>
- <https://www.periscopedata.com/amazon-redshift-guide>

Benchmarks

- <https://amplab.cs.berkeley.edu/benchmark/>
- [https://www.flydata.com/blog/with-amazon-redshift\(ssd\)-querying-a-tb-of-data-took-less-than-10-seconds/](https://www.flydata.com/blog/with-amazon-redshift(ssd)-querying-a-tb-of-data-took-less-than-10-seconds/)
- <https://www.flydata.com/blog/hive-and-redshift-a-brief-comparison/>

Thank you!