

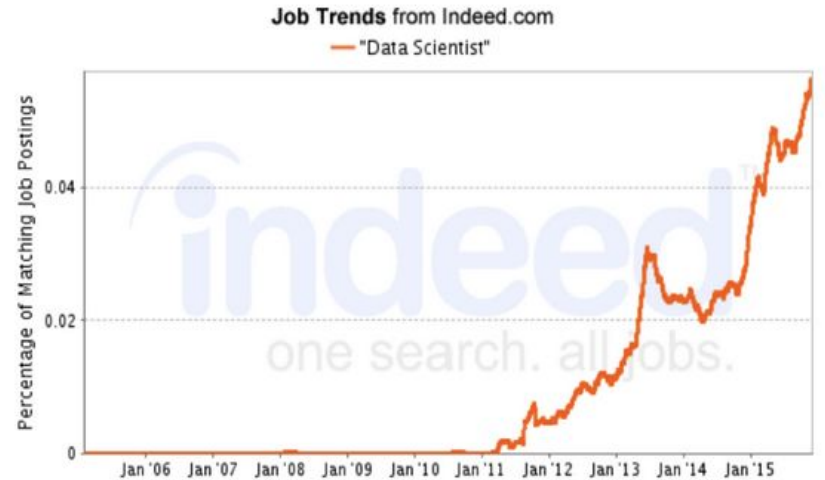
Services for data science at AWS

Luis Sosa, Paul Almasan



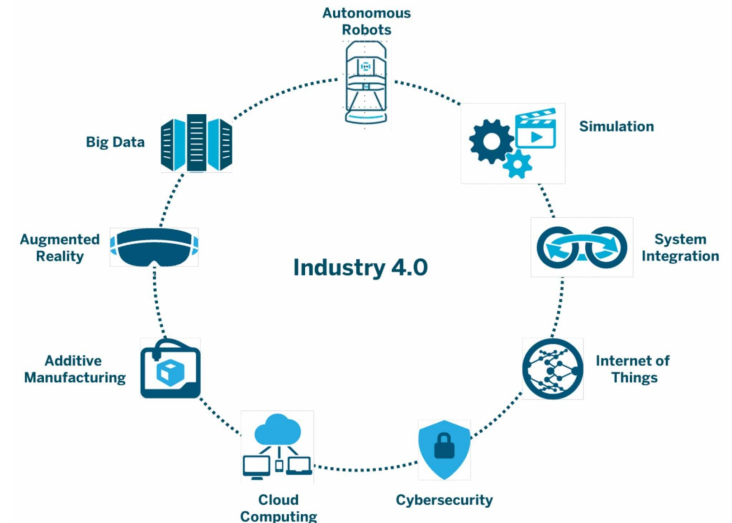
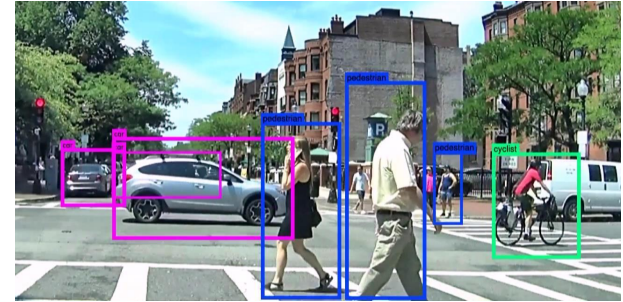
Needs of data science

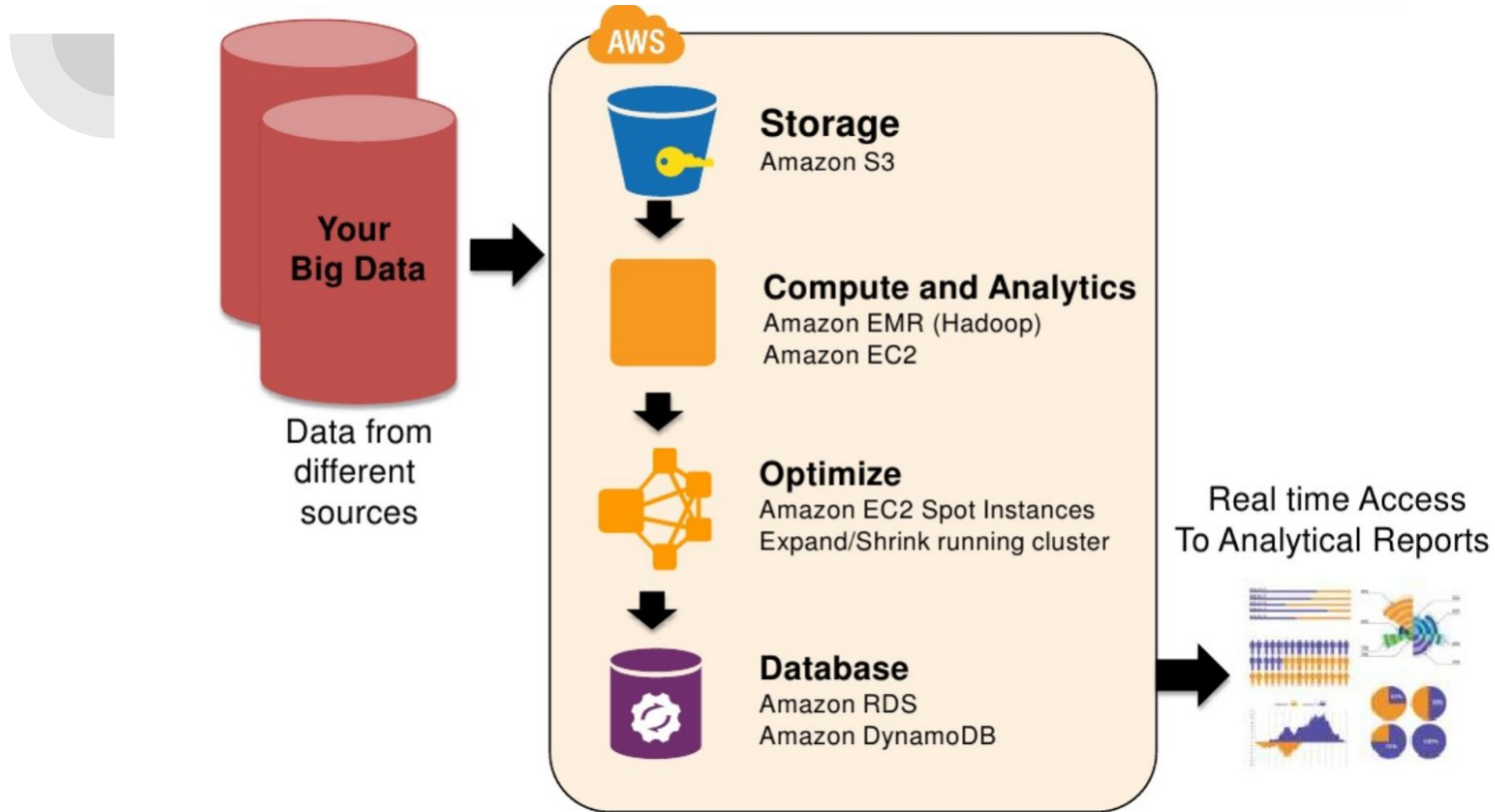
- **Extract knowledge** or insights from large amounts of data
 - Data extracted can be either structured or unstructured
- Data science tries to solve many issues within individual sectors and the economy at large



Applications

- Image recognition
- Natural Language processing
 - Search engines, text editor, information extraction
- Recommendation systems
- Industrial process optimization (i.e. Industry 4.0)







DynamoDB

- **NoSQL** database service
- Flexibility to design the optimal architecture for an application
- **Auto Scaling** parameter
- **Trigger feature** integrated with AWS Lambda
 - Code actions based on updates on items in the table
- DynamoDB has a **fault tolerance system** that replicates data across three data centers in a region



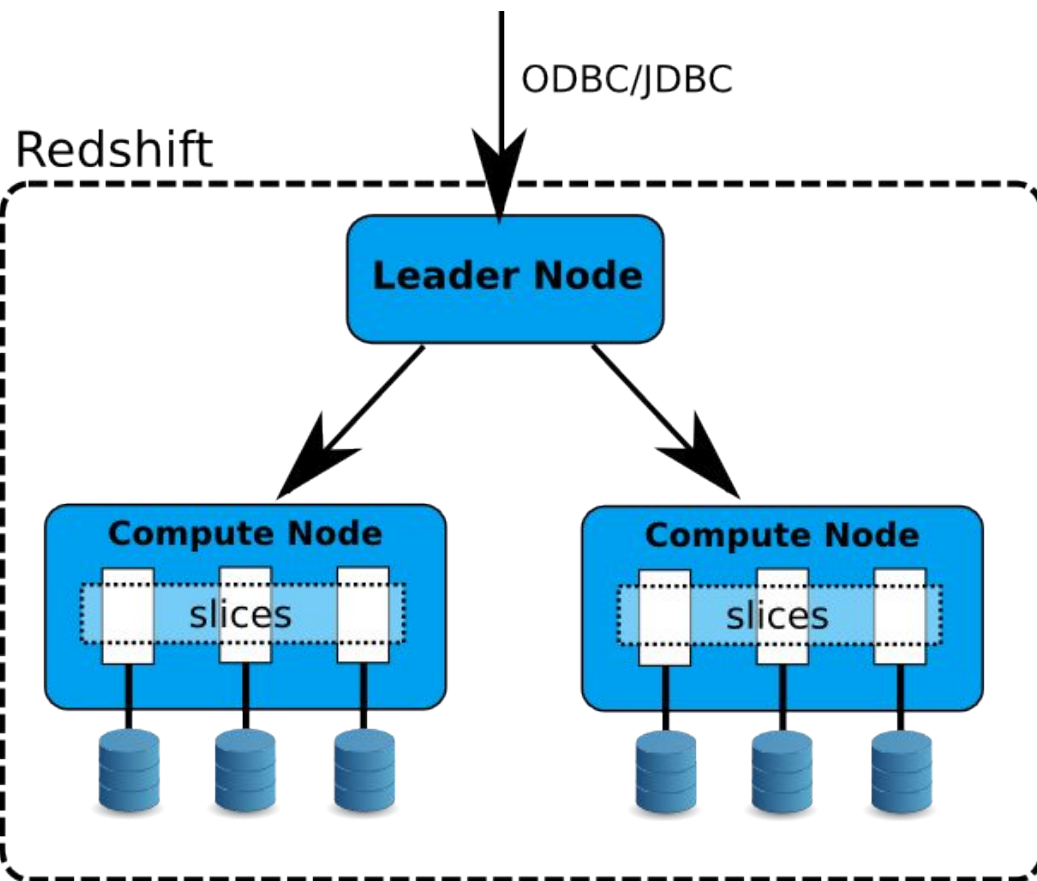
Simple Storage Service

- **Object storage service**
 - Each object is treated as a file and having an ID number
 - Up to **5 terabytes of data**
- A developer can access an object by using the **ID number and via a rest API**
- Designed to work with **online applications** and some libraries or APIs makes it easier to work with (e.g. **Boto3**)
- S3 is designed to deliver **99.999999999% of durability** and it has two storage classes: S3 Standard and S3 Infrequent Access



RedShift

- Data **warehouse service**
- Optimized for data sets ranging from a few **gigabytes to petabytes**
- Data is distributed on different nodes (servers) connected to a **cluster**
 - **Massively parallel processing architecture** to parallelize and distribute SQL operations
 - Query, they run in parallel on all the nodes
- If a **node fails**, is detected automatically and replaced





Elastic Map Reduce

- High **distributed framework** to process and store data in a cost effective manner
 - **Apache Hadoop** to distribute the processing and the data across a cluster of Amazon EC2 instances
- Cluster can have three types of nodes:
 - **Master** responsible of managing the cluster and distribute the workloads to core and task nodes
 - **Core nodes** that processes the tasks and stores data and the
 - **Task nodes** that can only run tasks

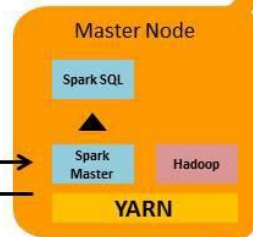


Amazon Simple Storage Service (Amazon S3)



Input data

Output data



Slave Node

Slave Node

Slave Node

Slave Node

Amazon Elastic MapReduce (Amazon EMR) Cluster

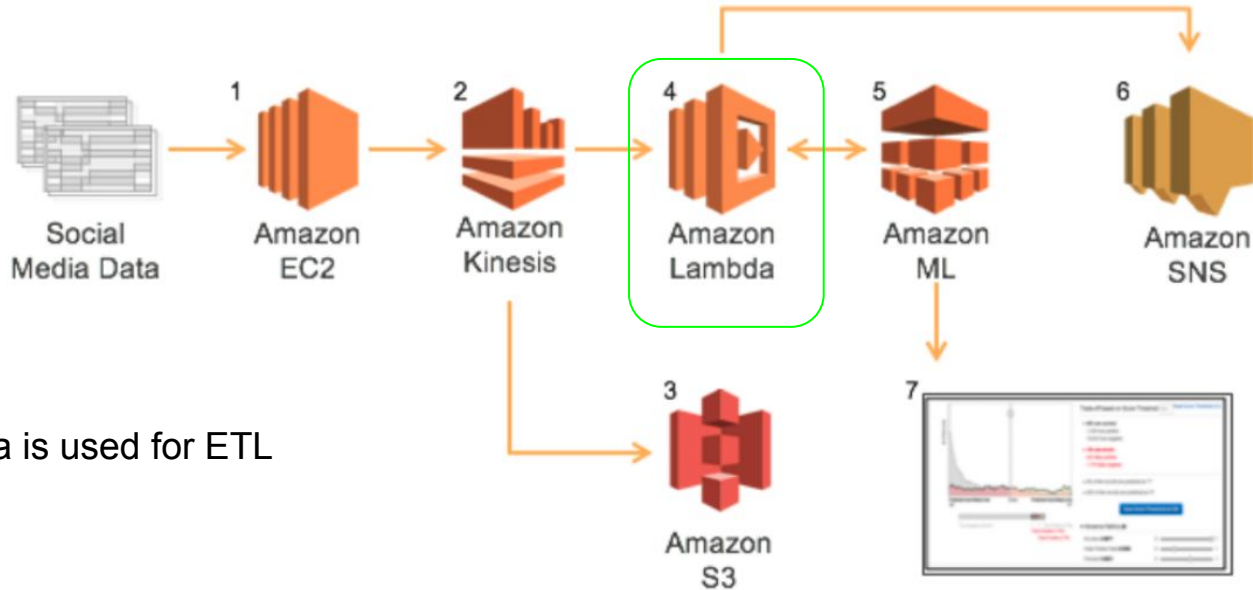


Lambda

- Execute an action as a response of a trigger
- ETL, send an email.
- Lambda supports different programming languages such as Java, Python, C#, etc.



Lambda



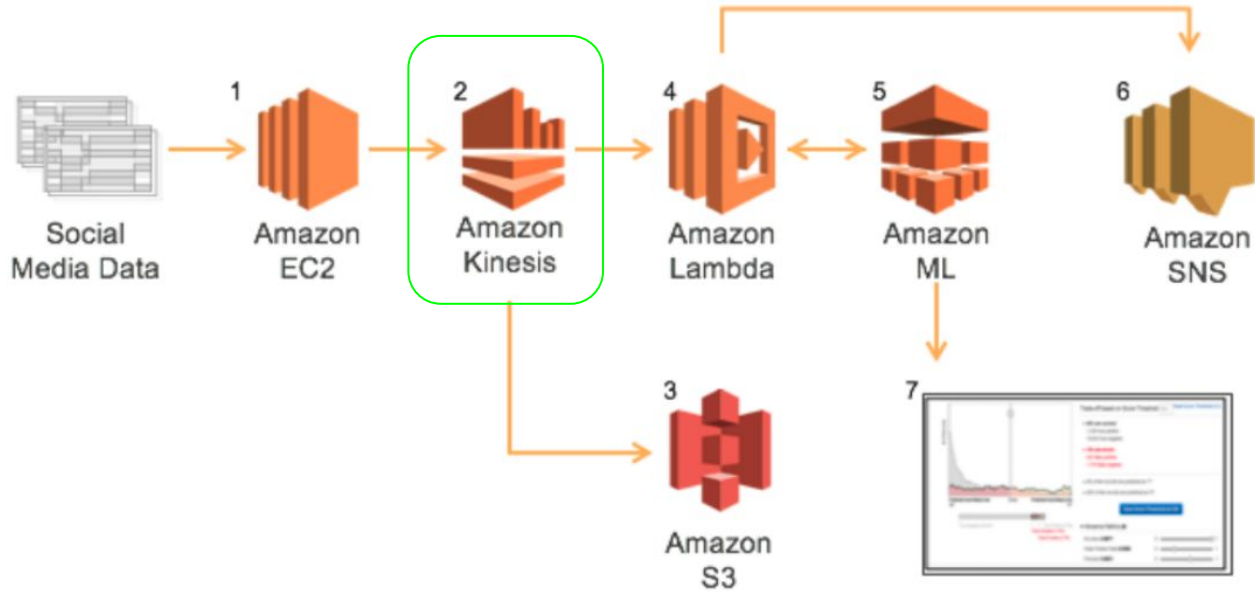
- Lambda is used for ETL



Kinesis stream and firehose

- It is a tool that gives the possibility to users to capture streaming of data in real time.
- Monitor uses behavior using logs.
- Netflix uses AWS to analyze billions of messages.
- Firehose: Moviles, apps.web apps, simplest, automatically scales, transfer data to S3
- Stream: Manual work to configure the service to handle the amount of data, data last for 24 hours, could be change to 7 days but additional cost

Kinesis stream and firehose



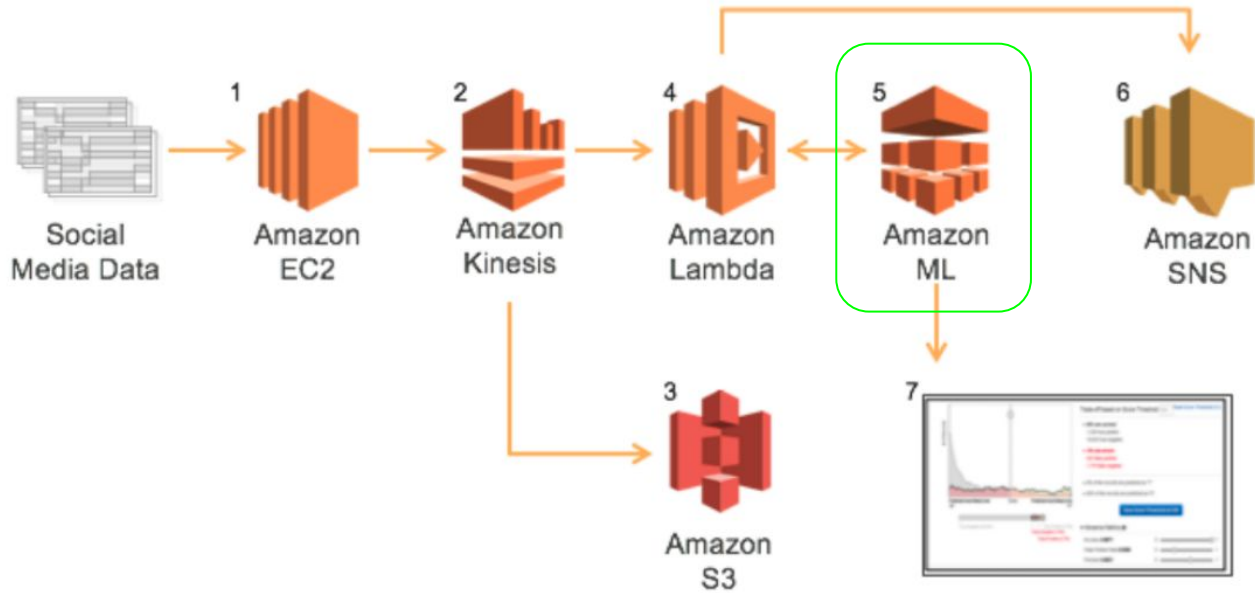
Sentiment Analysis of Social Media



Machine Learning

- Amazon machine learning is a server that provides to user with the tools to start using machine learning algorithms.
- Users can build models using the wizards and other visual tools.
- Upload file using wizard.
- Binary classification model, multiclass classification model, regression models.
- Multiples options to upload input for testing (batch or manual entry)
- Users don't have to worry about size of input data, (handle file of 100GB)
- It is easy to add a data source (just upload data to s3)

Machine learning



Sentiment Analysis of Social Media






QuickSight

- It is a service cloud-powered for business analytic that make it easy to build visualizations as well as create ad-hoc analysis
- This product has the capability to connect to a considerable amount of possible source of information.
- Graphs: Horizontal bar chart, Vertical Bar chart, Stacked bar chart, Line chart, Area line chart, Pivot table, Tables, Scatter plot, Tree map, Pie chart, Heat map, KPI (Key performance indicator), Clustered bar combo bar, points on map.



QuickSight (data source)


 QuickSight


 Ireland  6996476


Data sets


31.5MB of SPICE used of 1GB in Ireland


FROM NEW DATA SOURCES


 Upload a file
(.csv, .tsv, .cif, .elf, .xlsx, .json)


 Salesforce
Connect to Salesforce


 S3 Analytics


 S3


 Athena


 RDS


 Redshift
Auto-discovered


 Redshift
Manual connect


 MySQL


 PostgreSQL


 SQL Server


 Aurora


 MariaDB


 Presto


 Spark


 Teradata
Provided by Teradata


 Snowflake


 AWS IoT Analytics

 Github

 Twitter

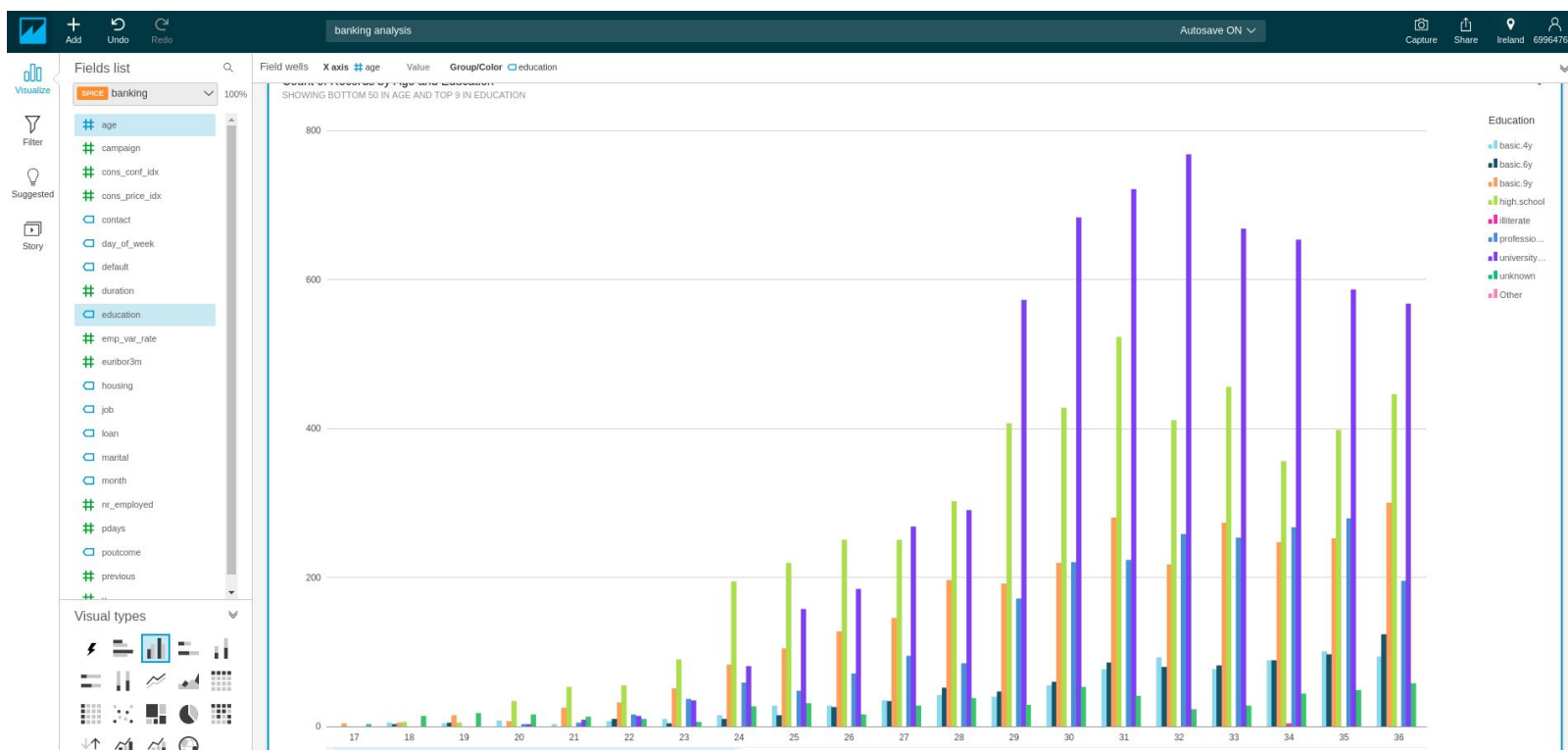
 Jira

 Service Now

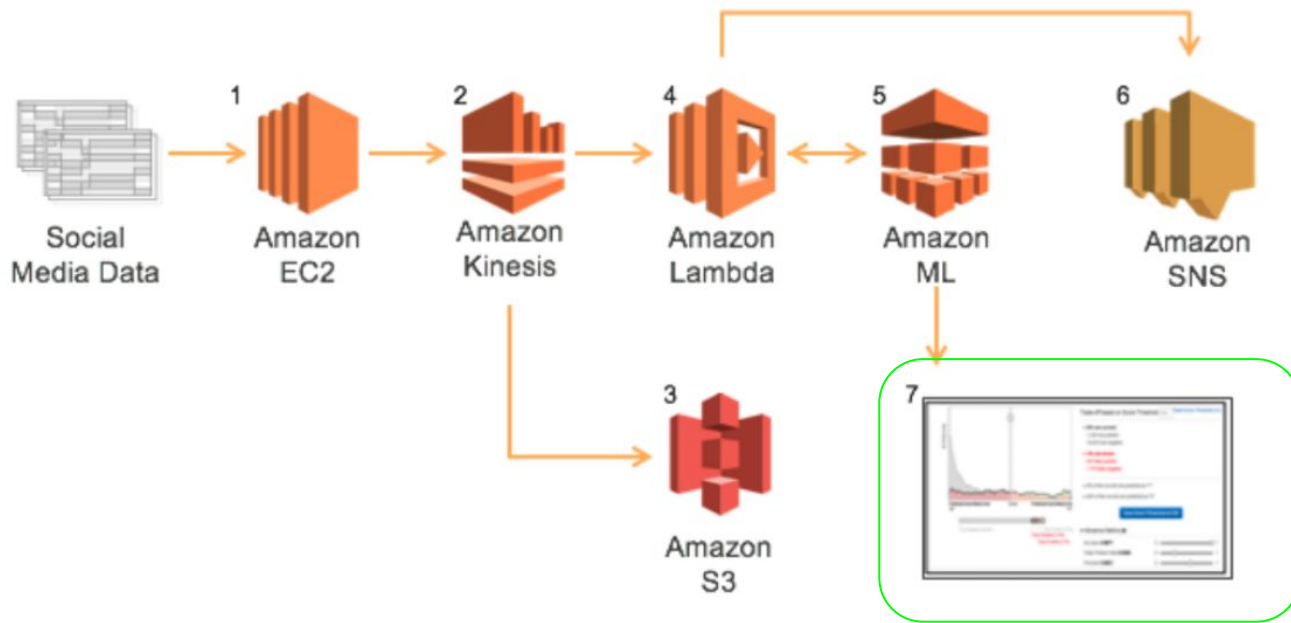
 Adobe Analytics

FROM EXISTING DATA SOURCES

QuickSight



QuickSight



Sentiment Analysis of Social Media



References

<https://github.com/paulalm94/CLOUD-COMPUTING-CLASS-2018/tree/master/Research-topic>