AWS

MACHINE LEARNING

CERTIFICATION

MODULE #2: EXPLORATORY DATA ANALYSIS (24% EXAM)

AWS ML CERTIFICATION EXAM DOMAINS

Domain	% of Examination
Domain 1: Data Engineering	20%
Domain 2: Exploratory Data Analysis	24%
Domain 3: Modeling	36%
Domain 4: Machine Learning Implementation and Operations	20%
TOTAL	100%

Source: https://d1.awsstatic.com/training-and-certification/docs-ml/AWS%20Certified%20Machine%20Learning%20-%20Specialty_Exam%20Guide%20(1).pdf

DOMAIN #2 OVERVIEW: WHERE ARE WE NOW?!

SECTION #5: JUPYTER NOTEBOOKS, SCIKIT LEARN, PYTHON PACKAGES, AND DISTRIBUTIONS

- Introduction
- Jupyter Notebooks and Scikit Learn
- Python Packages (Pandas, Numpy, Matplotlib and Seaborn
- Distributions (Normal, Standard, Poisson, Bernoulli)
- Time Series

WE ARE HERE!

SECTION #6: AMAZON ATHENA, QUICKSIGHT AND ELASTIC MAP REDUCE

- Amazon Athena Features
- Amazon Athena Deep Dive (Security, Cost, and glue integration)
- Amazon QuickSight Features
- Amazon QuickSight (integration with AWS services)
- Amazon QuickSight ML insights and Use Cases
- Elastic Map Reduce (EMR)
- Apache Hadoop with EMR
- Apache Spark with EMR

DOMAIN #1 OVERVIEW:

SECTION #7: FEATURE ENGINEERING

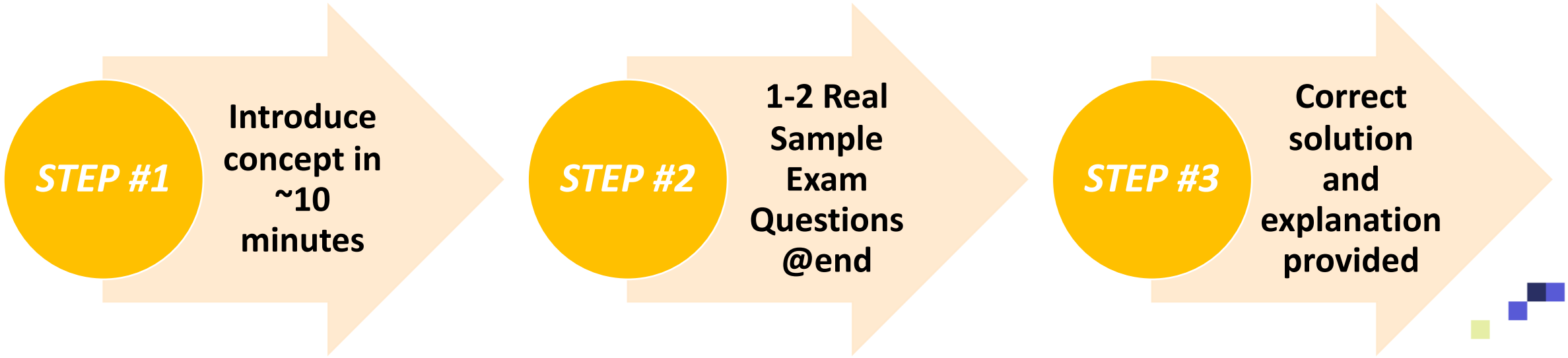
- Introduction to Feature Engineering
  - Amazon SageMaker GroundTruth
  - Feature Selection
  - Scaling
  - Imputation
  - Outliers
  - One Hot Encoding
  - Binning
  - Log Transformation
  - Shuffling, Feature Splitting, Unbalanced Datasets
  - Text Feature Engineering overview
  - Bag of words, punctuation, and dates (easy ones!)
  - Term Frequency Inverse Document Frequency (TF-IDF)
  - N-Grams (Unigram vs. Bigram vs. Trigram)
  - Orthogonal Sparse Bigram (OSB)
  - Cartesian Product Transformation

LECTURE DESIGN

- We know how hard it is to study for an exam especially if you have a busy schedule.
- This course is designed to be extremely on point and optimized to pass the exam.

No boring content. Zero unnecessary information.

- Here's the lecture structure that we will follow:



RECALL OUR MINI CHALLENGE AND PRIZE!

- For those of you who will successfully complete the entire section and watch the videos till the end, they will receive a valuable prize!

https://pixabay.com/illustrations/trophy-win-prize-transparent-1414791/

AMAZON ATHENA – PART #1

AMAZON ATHENA: FEATURES

- The Amazon Athena is a flexible, cost-effective query service.
- Amazon Athena is used for data analysis simply by accessing the data available in Amazon S3 using standard SQL requests. So if you know basic SQL, you can start using Athena now to analyze large scale datasets!
- Athena eliminates the need to have complex and expensive ETL* jobs to analyze your data.
- Athena is serverless which makes it extremely easy to use.
- Athena is very fast, results are retrieved within seconds.
- It is very cost effective, you only pay per queries you choose to run.

- *ETL stands for extract, transform, load.*
- *ETL combines multiple functions to extract data from one database and put it in another one.*

Photo Credit: https://publicdomainvectors.org/en/free-clipart/Athena-in-black-and-white/82198.html

AMAZON ATHENA: FEATURES

- There is no infrastructure management overheads.
- Here's how Athena works:
  1. Point to S3 bucket where your data is located
  2. Define a schema
  3. Using SQL, start running queries to access the data
- Athena is easily integrated with AWS Glue data catalogue, this integration allows you to:
  - Develop a unified metadata repository with multiple services.
  - Crawl multiple data sources to extract the schemas.
  - Populate Catalog with new tables and maintain schema versioning.
  - Use Glue's fully-managed ETL services to convert data into columnar format for improved performance and cost effectiveness.

AMAZON ATHENA: FEATURES

- Amazon Athena is extremely fast, it is capable for executing multiple queries using several compute resources across multiple machines.
- Since Athena relies on data available on Amazon S3, this makes the data highly available and durable.
- Amazon Athena uses Presto with ANSI SQL support.
- Athena supports several formats such as:
  - CSV
  - JSON
  - ORC
  - Avro
  - Parquet
- Athena works great with both:
  - quick querying
  - Complex analysis

Photo Credit: https://commons.wikimedia.org/wiki/File:Data-transfer.svg

AMAZON ATHENA: FEATURES AND BENEFITS

FAST DATA QUERY

- Serverless
- No ETL
- No server management and zero overheads.
- pay per query so very cost optimized
- Access all your data buckets in S3 with zero ETL overheads.

COST EFFECTIVE

- You only pay per queries.

VERY HIGH SPEED

- Extremely fast so obtain results in seconds.
- No need to worry about having powerful compute resources to obtain fast speed. Athena manages that and executes queries in parallel.

STANDARD/OPEN

- Built on Presto.
- Runs standard SQL.

AMAZON ATHENA: FEATURES AND BENEFITS

DURABILITY & HIGH AVAILABILTY (S3 FOR DATA STORE)

- Amazon Athena leverages multiple compute resources across several facilities so this makes it highly available.
- Athena uses Amazon S3 so the data becomes available, secure and durable.
- Athena inherits the durability of Amazon S3 which is set at 99.999999999% of objects. Your data is redundantly stored across multiple facilities and multiple devices in each facility.

VERY SECURE

- Athena is secure because it leverages:

1. AWS Identity and Access Management (IAM) policies
2. Access control lists (ACLs)
3. Amazon S3 bucket policies

SEAMLESS INTEGRATION WITH OTHER AWS SERVICES (GLUE)

- Amazon Athena can be seamlessly and effectively integrated with AWS Glue.

AMAZON ATHENA – PART #2

ATHENA SECURITY

- Athena is secure because it leverages:
  - AWS Identity and Access Management (IAM) policies
  - Access control lists (ACLs)
  - Amazon S3 bucket policies
- Using S3 bucket polices, you can allow or prevent users from querying it using Athena.
- Using IAM policies, you can grant/deprive IAM users access to various buckets in amazon S3.
- Athena allows encryption for both client and server sides.
- Athena enable users to query encrypted data stored in Amazon S3.
- It also allows users to write encrypted results back to S3 buckets.
- Transport Layer Security (TLS) encrypts in transit data between Athena and Amazon S3.

Photo Credit: https://www.flickr.com/photos/mikemacmarketing/35366000233

ATHENA: COST MODEL

- Users are only charged for the queries they run.
- Users are charged based on data size scanned during each query.
- How do users reduce cost and improve performance?
  - Compressing, partitioning, or converting the data to a columnar format.
  - This dramatically reduces the time and resources Athena requires to scan and execute queries.
  - This could result in 30% to 90% cost reduction.
- Users are charged $5 per terabyte scanned (rounded up to nearest megabyte, with a 10MB minimum per query).
- If you cancel a query, users will be charged based on the amount of data scanned.
- Athena supports Apache ORC and Apache Parquet.
- Note: Amazon S3 and Glue have separate charges.

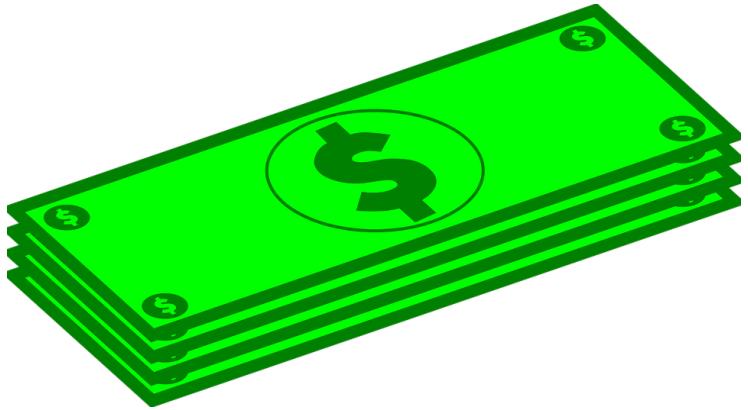


Photo Credit: https://pixabay.com/vectors/cost-currency-dollars-four-green-151072/

GLUE AND ATHENA

AMAZON S3

AWS GLUE CRAWLERS

AWS GLUE DATA CATALOGUE

AMAZON ATHENA

AMAZON QUICKSIGHT

AMAZON REDSHIFT

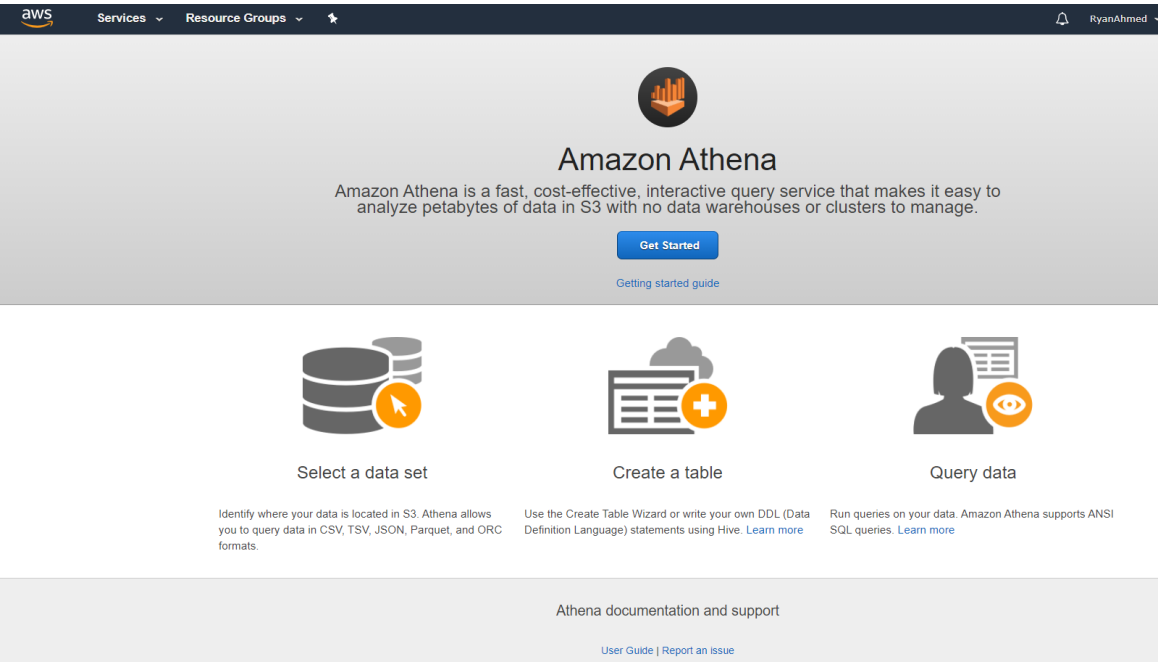
Setup a crawler to scan the S3 data periodically and generate Glue data catalogue

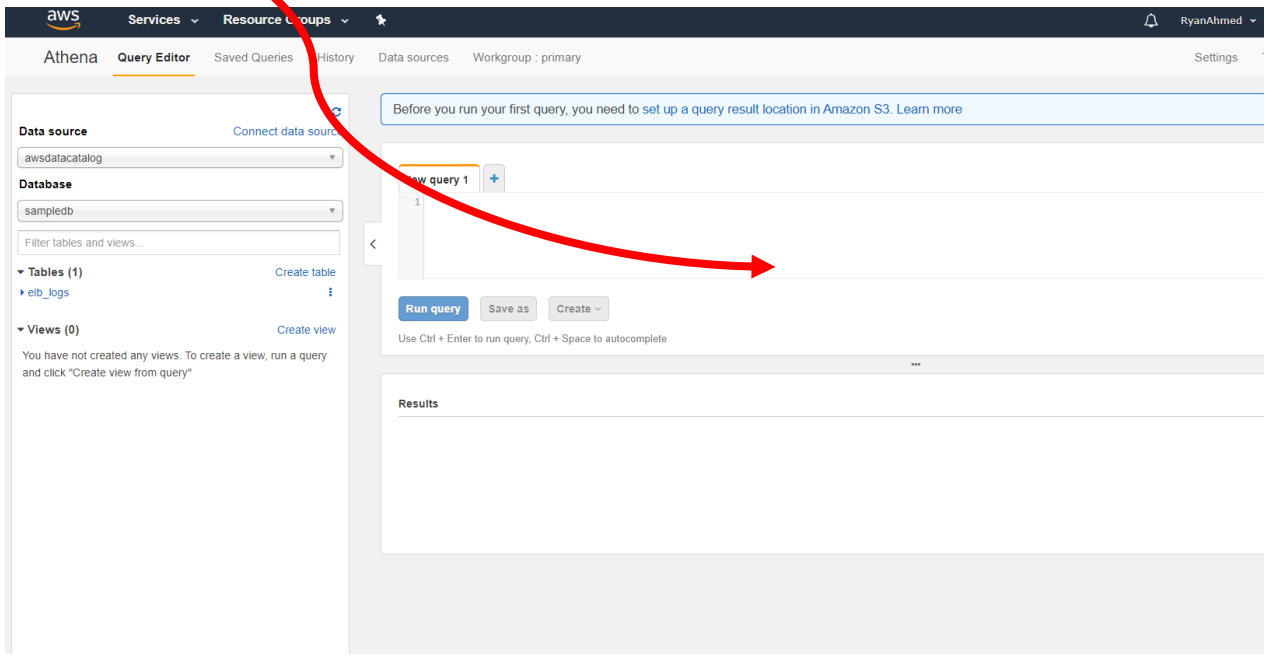
Once the AWS Glue data catalogue is available, Amazon Athena and Redshift can query the data and view in Quicksight

- *Athena and AWS Glue Data Catalog work seamlessly together, AWS Glue can be used to create databases and tables (schema) so that Athena can use it to query the data.*

ATHENA: IN ACTION

INSERT QUERY HERE





ATHENA VS. REDSHIFT SPECTRUM

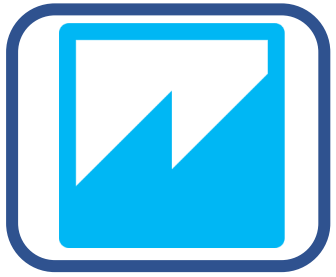
- Do you remember Amazon Redshift Spectrum?
- Redshift spectrum was used to generate queries as well directly into S3 which seem similar to what Athena does!

ATHENA 	REDSHIFT SPECTRUM 
Send Queries directly to Amazon S3	**Send Queries directly to Amazon S3**
Designed for easy ad-hoc queries into S3	**Designed for users of Redshift**
Does not require redshift clusters	**You will require redshift clusters.**

AMAZON QUICKSIGHT – PART #1

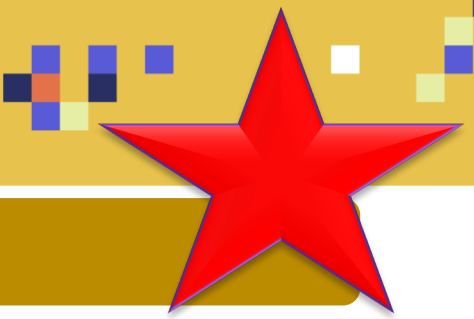
AMAZON QUICKSIGHT

- Amazon QuickSight is cloud business intelligence (BI) tool that is used to visually share data across the entire organization.
- QuickSight is a fully managed service, and therefore it is extremely easy to use.
- It allow anyone to easily develop interactive dashboards along with ML Insights (will be introduced in details shortly!).
- Users can view the dashboard from anywhere on any device.
- Users can also embed the dashboard into any application or websites.
- Quicksight follows a per-per-use model so you only pay for what you use.



AMAZON QUICKSIGHT

AMAZON QUICKSIGHT FEATURES



COST EFFECTIVE, PAY PER USE

- QuickSight's follows a pay-per-session model.
- Charges are applied only when users access dashboards or reports.
- No annual subscription and zero upfront costs.
- Zero charges for users who are inactive.

HIGHLY SCALABLE, SCALE FROM 10 TO 10,000 USERS

- Serverless architecture, zero servers to manage, and no installation or setup required
- No need for capacity planning or any infrastructure cost.
- QuickSight scales automatically to 10,000 users.

ALLOW FOR EASY DATA ANALYTICS EMBEDDING

- Quicksight allows for embedding of dashboards and charts into applications.
- QuickSight visuals can be securely embedded in the application with authentication and powerful APIs.

END-TO-END BUSINESS INTELLIGENCE SOLUTION

- QuickSight works seamlessly with other AWS services available on the cloud such as RedShift, S3, Athena, Aurora, RDS, IAM, CloudTrail, Cloud Directory.
- This allows for building an end to end complete BI solution across any organization.

AMAZON QUICKSIGHT: HOW IT WORKS?

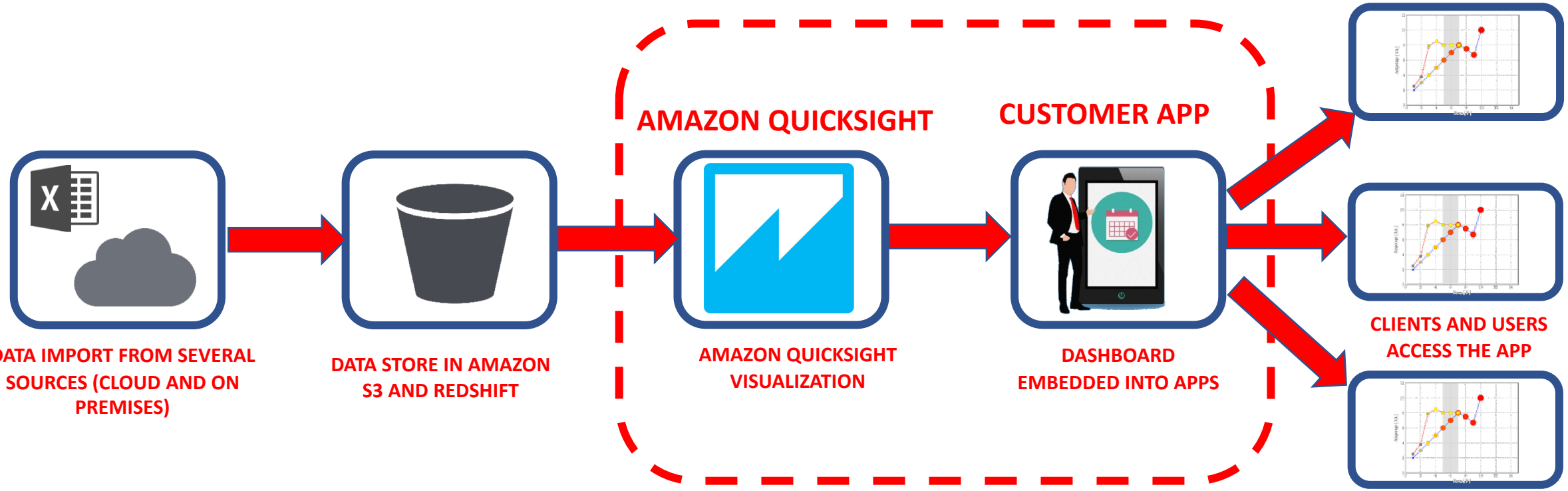


Photo Credit: https://commons.wikimedia.org/wiki/File:AWS_Simple_Icons_Storage_Amazon_S3.svg
Photo Credit: https://commons.wikimedia.org/wiki/File:Document-open.svg
Photo Credit: http://pgfplots.net/tikz/examples/plot-markers/
Photo Credit: https://commons.wikimedia.org/wiki/File:Microsoft_Excel_2013_logo.svg
Photo Credit: https://pixabay.com/vectors/cloud-cloud-computing-3331240/
Photo Credit: https://pxhere.com/en/photo/1439573

AMAZON QUICKSIGHT: DATA SOURCES

There are several sources of data available for consumption and visualization by Amazon Quicksight:

- Amazon Redshift
- Athena
- S3 or on-premises files in the following formats:
  - Excel
  - CSV
  - TSV
- EC2-hosted databases
- Aurora/RDS

- ***Let's take a look at an integration example with Amazon Athena!***

AMAZON QUICKSIGHT + ATHENA + S3?

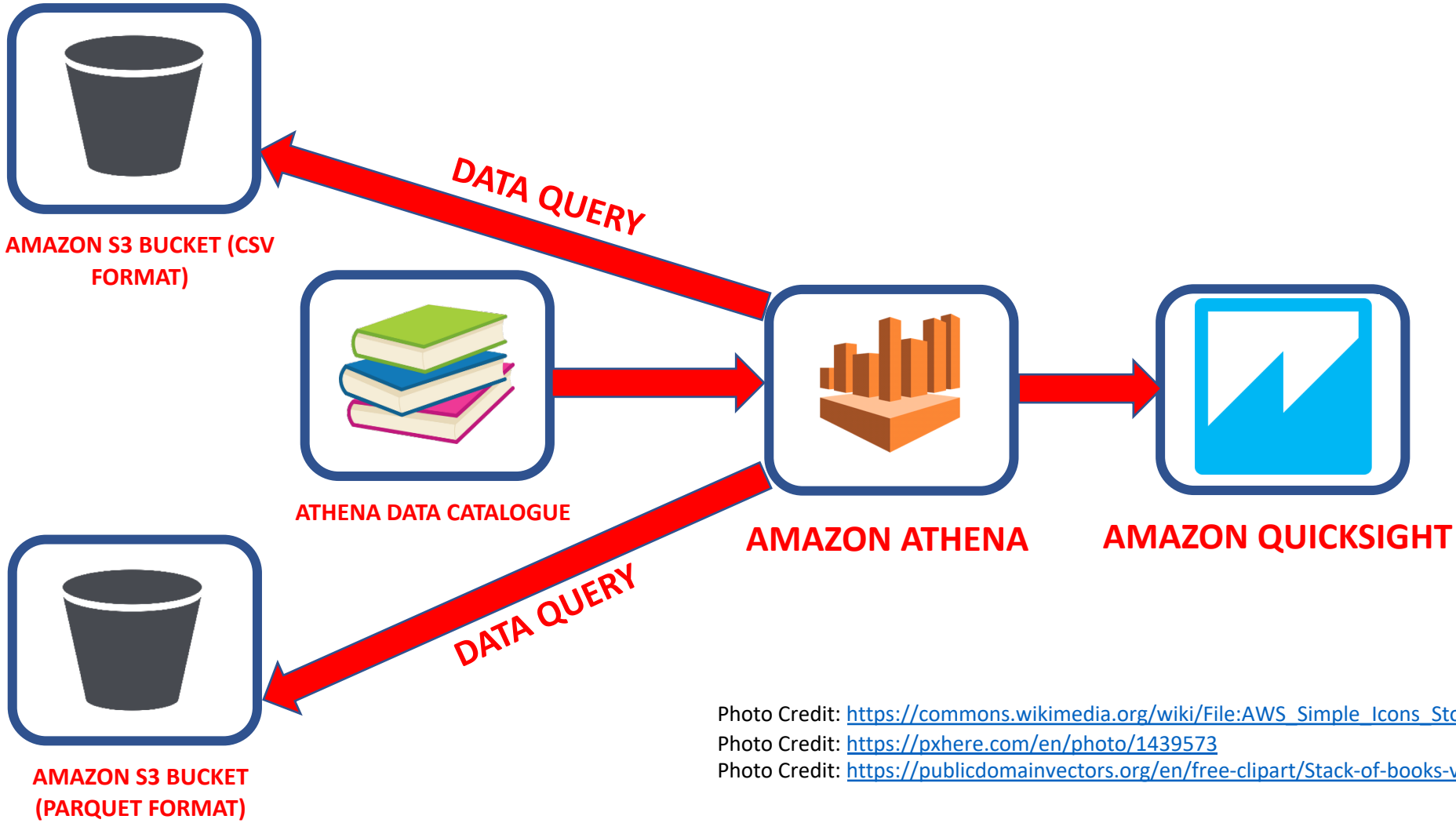
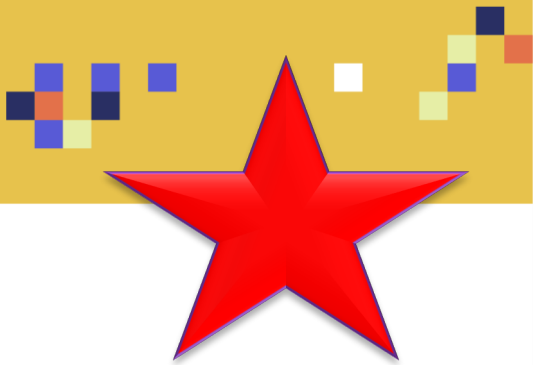


Photo Credit: https://commons.wikimedia.org/wiki/File:AWS_Simple_Icons_Storage_Amazon_S3.svg
Photo Credit: https://pxhere.com/en/photo/1439573
Photo Credit: https://publicdomainvectors.org/en/free-clipart/Stack-of-books-vector-clip-art/75624.html

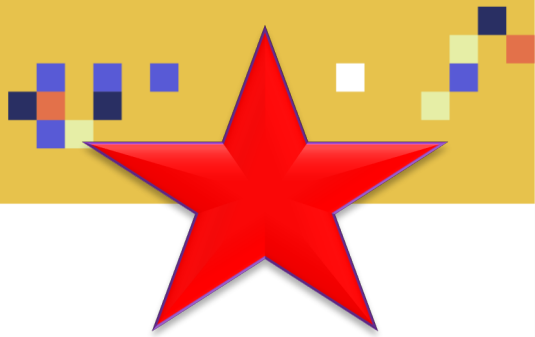
AMAZON QUICKSIGHT – PART #2

AMAZON QUICKSIGHT: SPICE



- SPICE is a fast, optimize, in-memory calculation engine for Amazon QuickSight.
- SPICE is Highly available and durable, and could be scaled to hundreds of thousands of users.
- SPICE can be used for fast, ad-hoc data visualization.
- SPICE stores the data in a system that enables fast access, data is saved until it is deleted by user.
- Once you have a QuickSight account, you can:
  - For paid users, automatically get 10 GB of SPICE capacity
  - Free users get 1 GB of SPICE capacity.
  - Purchase additional capacity.
- Instead of sending direct queries to the database, you can import data into SPICE for great performance improvements.
- If you do not use data in SPICE, you can simply delete unused data.

AMAZON QUICKSIGHT: USE CASES



PACKAGED DATA PRODUCTS (SELL DATA IN A PACKAGED FORMAT)

- QuickSight can enable enterprises to share reports and dashboards with customers.
- Data Monetization by offering a packaged product.

IMPROVE APPS BY OFFERING ANALYTICS

- Improve app experience by offering customers rich dashboards/reports while integrating Quicksight ML Insights features.

INTEGRATE DATA INTO WORKFLOWS

- By embedding Quicksight's rich dashboards/reports into portals and company sites.

"Amazon QuickSight will allow us to quickly build fast, interactive dashboards that will seamlessly integrate with our Next Gen Stats applications. With the Amazon QuickSight Readers and pay-per-session pricing, we are able to extend these secure, customized and easy to use dashboards for each Club without having to provision servers or manage infrastructure – all while only paying for actual usage. We love the direction, and look forward to expanding use of Amazon QuickSight."
Matt Swensson, VP Emerging Products - NFL

Source: https://aws.amazon.com/quicksight/features-embedding/?nc=sn&loc=3

AMAZON QUICKSIGHTS: EMBEDDED ANALYTICS

- QuickSight embedding allows for easy integration of analytics in apps and websites.
- It is cost effective and fully managed service.

**SDK is a collection of software used for app development targeted at an operating system such as Windows 10 SDK, Mac OS X SDK, and iPhone SDK*

PROVIDE A COMPLETE ANALYTICS EXPERIENCE

- Allows for creating amazing, modern dashboards that uses QuickSight's visualization and analytics capabilities(ML Insights) and Auto-Narratives.

RICH APIs AND SDK*

- Powered by JavaScript SDK, it could allow for integration between the application and the embedded dashboards powered by QuickSight.

A SCALABLE, GLOBAL PLATFORM

- QuickSight is a fully managed, secure service.
- Supports 10 languages with exceptional end-to-end data security.
- Fast performance powered by SPICE.

EASY TO DEVELOP AND MAINTAIN

- Easily develop dashboard templates that you can embed into your application.

AMAZON QUICKSIGHT: ML INSIGHTS

- Amazon QuickSights offer an integrated ML insights features.
- ML insights offer the following:
  1. Find data insights:
     - QuickSight's ML offers an anomaly detection tool.
     - The tool can perform anomaly detection and notify management.
  2. Forecasting:
     - Quicksight's ML tool can perform accurate forecasting to predict critical business metrics.
  3. Generate Auto-Narratives:
     - Powered by NLP, QuickSight's ML can offer narratives and tell a story!

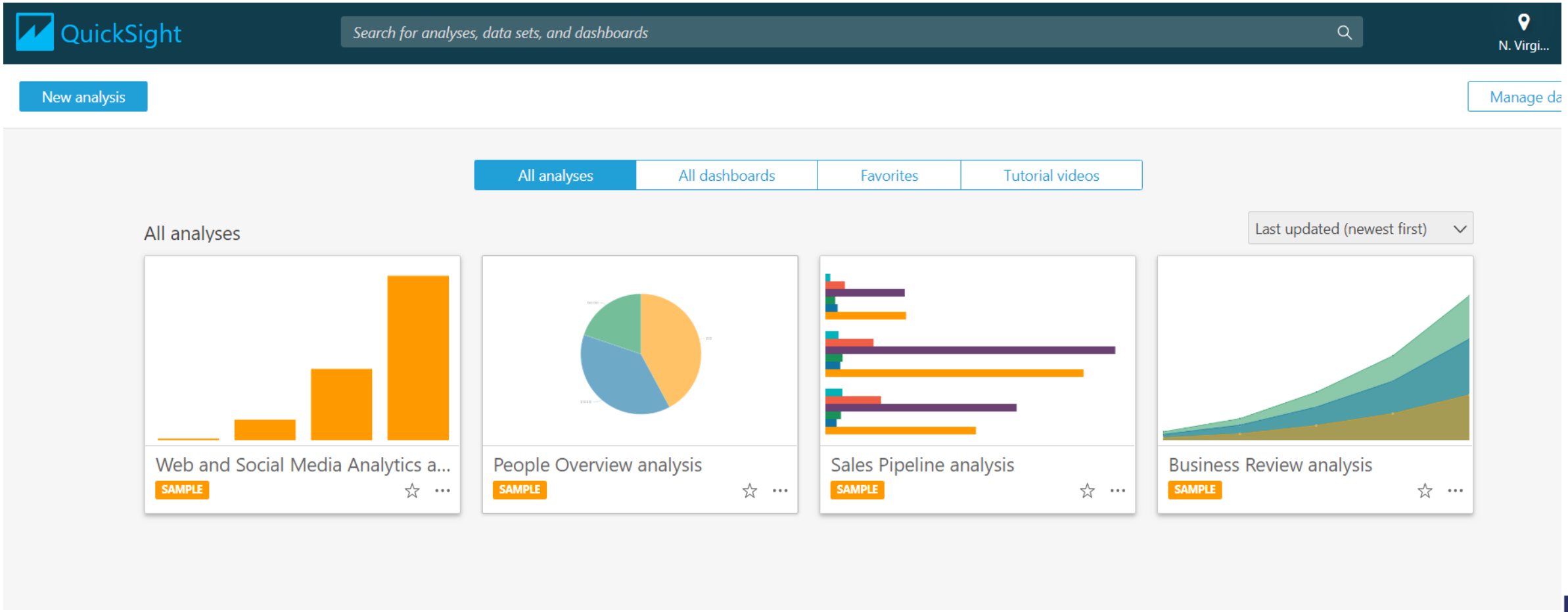
Photo Credit: https://pixabay.com/illustrations/mindset-programmer-machine-learning-3536805/

AMAZON QUICKSIGHT: PRICING

- Check out pricing here: https://aws.amazon.com/quicksight/pricing/?nc=sn&loc=4
- The pricing model varies if you are an author or reader.
- For authors:
  - Annual subscription
    - Standard: $9/user/month
    - Enterprise: $18/user/month
  - Additional SPICE capacity more than 10GB
    - $0.25 (standard) / GB / month
    - $0.38 (enterprise) / GB / month
- For the ML-powered Anomaly Detection, there is a pricing model.

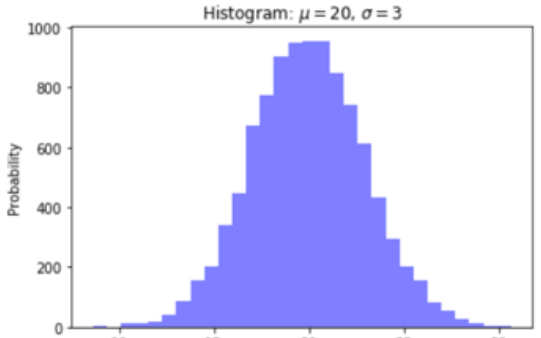
VOLUME TIER	$ PER THOUSAND METRICS PROCESSED PER MONTH
First 1,000,000 metrics processed	$0.50
Next 9,000,000 metrics processed	$0.25
Next 90,000,000 metrics processed	$0.10
> 100,000,000 metrics processed	$0.05

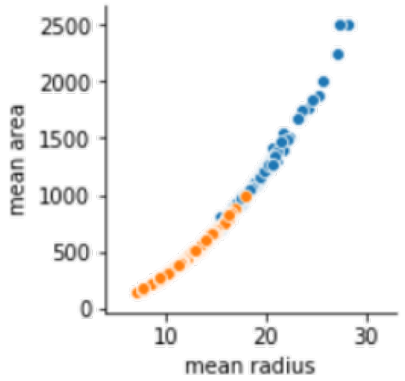
WHAT DOES QUICKSIGHT LOOK LIKE?

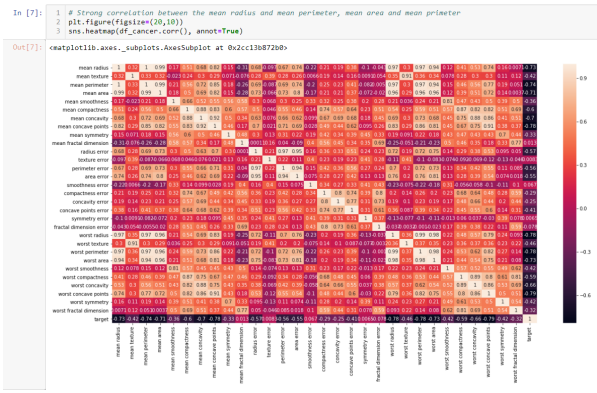


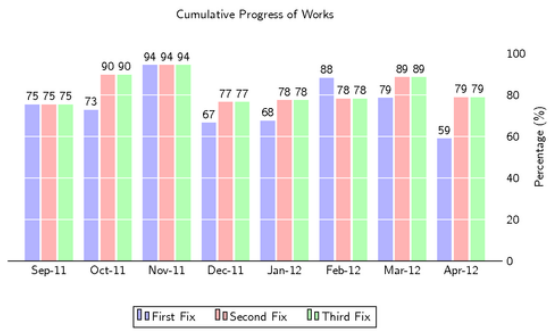
[Sample Dashboard #1: https://aws.amazon.com/blogs/big-data/embed-interactive-dashboards-in-your-application-with-amazon-quicksight/](https://aws.amazon.com/blogs/big-data/embed-interactive-dashboards-in-your-application-with-amazon-quicksight/)
[Sample Dashboard #2: https://aws.amazon.com/blogs/big-data/amazon-quicksight-updates-multiple-sheets-in-dashboards-axis-label-orientation-options-and-more/](https://aws.amazon.com/blogs/big-data/amazon-quicksight-updates-multiple-sheets-in-dashboards-axis-label-orientation-options-and-more/)

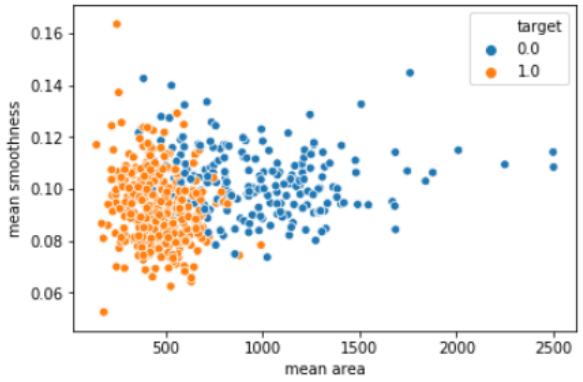
WHAT VISUALIZATIONS CAN WE MAKE?

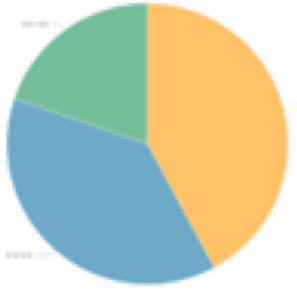












	first	last	email	postal	gender	dollar
0	Joseph	Patton	daafeja@boh.jm	M6U 5U7	Male	$2,629.13
1	Noah	Moran	guutodi@bigwoc.kw	K2D 4M9	Male	$8,626.96
2	Nina	Keller	azikez@gahew.mr	S1T 4E6	Male	$9,072.02

- **Bar Charts:** Comparison
- **Line graphs:** trends with time
- **Scatter plots/heat maps:** correlation
- **Pie chart:** aggregation
- **Pivot tables:** tabular data

Photo Credit: http://pgfplots.net/tikz/examples/multi-series-bar-chart/

QUICKSIGHT SECURITY

- Amazon QuickSight offers a secure service to allow users to access interactive dashboards from any device.
- Amazon QuickSight offers multiple security features such as:
  - Role-based access control
  - Microsoft Active Directory integration
  - AWS CloudTrail auditing
  - Single sign-on using AWS Identity and Access Management (IAM) and third-party solutions
  - Private VPC subnets (Elastic Network Interface, AWS Direct Connet)
  - Data backup
  - Multifactor authentication on the user account
  - Row level security
- Amazon QuickSight can also support FedRAMP, HIPAA, PCI DSS, ISO, and SOC compliance.

Photo Credit: https://pixabay.com/vectors/lock-padlock-green-locked-33495/

QUICKSIGHT SECURITY

- There are two levels of security of the cloud and security in the cloud:
  - Security of the cloud:
    - AWS is responsible for protecting the infrastructure that runs AWS services in the AWS Cloud.
    - Security levels are in compliance and being regularly tested by third-party auditors.
  - Security in the cloud:
    - Enterprises are responsible for the data sensitivity, the requirements, and applicable laws and regulations.

Photo Credit: https://pixabay.com/vectors/lock-padlock-green-locked-33495/

ELASTIC MAP REDUCE (EMR) – PART #1

WHAT IS EMR?

- EMR Stands for Elastic Map Reduce.
- Amazon EMR is big data platform that allows for data processing in a fast, easy and cost optimized way.
- Leverages Apache Spark, Apache Hive, Apache HBase, Apache Flink, and Presto.
- EMR empower developers to use:
  - Short Single-purpose clusters that are scalable based on demand.
  - Long term highly available clusters.
- EMR allows for:
  - Dynamic scalability using Amazon EC2
  - Storage of Amazon S3
- Using Jupyter-based EMR Notebooks, developers can work with data anywhere in AWS such as Amazon S3, Amazon DynamoDB, and Amazon Redshift.
- Works great with machine learning, data transformations (ETL), and deep learning.

WHAT IS EMR?

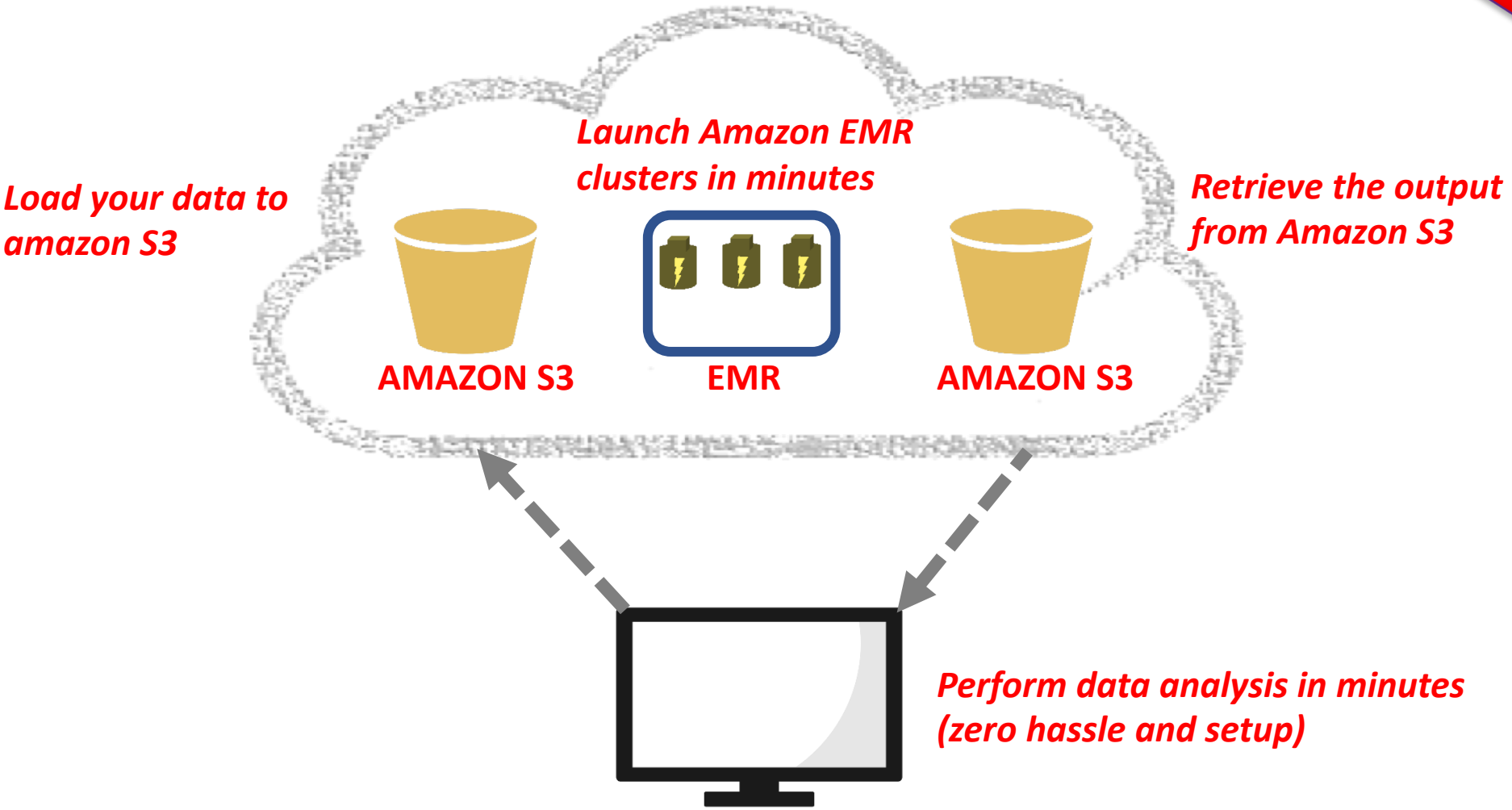


Photo Credit: https://publicdomainvectors.org/en/free-clipart/Grey-cloud-icon-vector-image/19199.html
Photo Credit: https://www.needpix.com/photo/824215/computer-pc-monitor-screen-computer-monitor-computer-screen-personal-computer-technology-laptop

EMR FEATURES – PART #1

EASE OF USE

- Developers could launch clusters in a very short period of time.
- Does not require any infrastructure setup or node provisioning.
- Works with serverless Jupyter notebooks (EMR Notebooks) so developers can analyze/visualize data easily.

COST EFFECTIVE

- EMR pricing model is effective, pay per instance per time. .
- Example: Launch multi-node EMR clusters and run applications such as Apache Spark, and Apache Hive and pay $0.15 per hour.
- Great savings can be achieved by leveraging Amazon EC2 Spot and Reserved Instances.

ELASTIC

- EMR gives you the flexibility to elastically change the compute and storage offering extremely cost effective way.
- EMR allows for instantiating whatever number of compute instances (you can increase/decrease at anytime) to work with massive amount of data at any scale.
- Auto Scaling feature allows for automatically managing cluster sizes based on utilization).

EMR FEATURES – PART #2

ENHANCED RELIABLILTY

- No need to waste time monitoring instances.
- Multiple master nodes are available to ensure high reliability.
- EMR automatically and periodically monitors instances, assesses their health and replace failed ones.
- EMR manages software updates resulting in minimum issues and less maintenance.

IMPROVED SECURITY

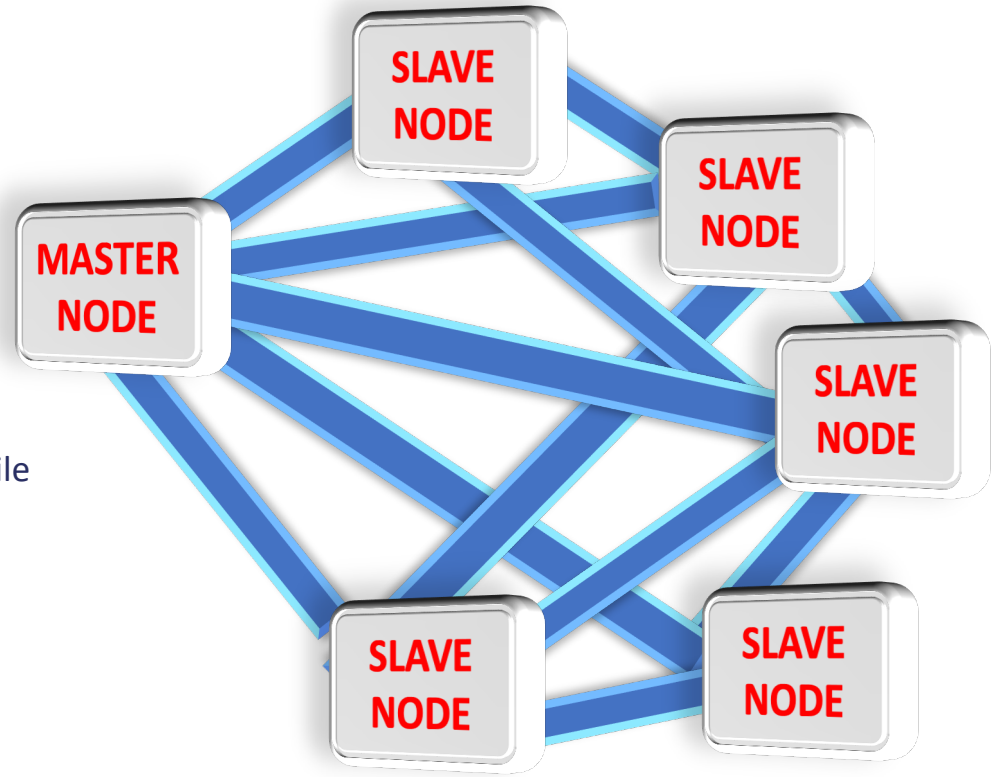
- EC2 firewall settings are automatically configured by EMR.
- Launches clusters Amazon Virtual Private Cloud (VPC)
- Server-side and client side encryption in S3 used with EMRFS (an object store for Hadoop on S3).
- AWS Key Management Service.
- Allows for In-transit and at-rest encryption
- Authentication with Kerberos

INCREASED FLEXIBLITY

- Increased flexible Cluster Control
- Developers have root access to instances which allow for customization with bootstrap actions.
- Launch EMR clusters with custom Amazon Linux AMIs, and reconfigure running clusters on the fly.

EMR CLUSTER

- A cluster consists of a group of Amazon Elastic Compute Cloud (Amazon EC2) instances.
- Instances are known as nodes.
- Several node type are available (runs different SW):
  - **Master node:**
    - Big boss! Node that manages the cluster.
    - Distribute data and tasks
    - Tracks status of tasks and monitors health cluster.
  - **Core node:**
    - A slave node
    - Runs tasks and store data in the Hadoop Distributed File System (HDFS) on the cluster.
  - **Task node:**
    - A slave node.
    - Optional
    - It only run tasks.
    - Spot instances.
    - No risk of data losses if task node is removed



"CLUSTER WITH ONE MASTER NODE AND FIVE SLAVE NODES"

EMR NOTES:

- EMR leverages several AWS services such as:
  - **Amazon EC2:** run instances (nodes) in the cluster
  - **Amazon S3:** data storage and data output.
  - **Amazon CloudWatch:** cluster performance monitoring and generating alarms
  - **IAM:** grant users permissions
  - **AWS CloudTrail:** audit requests made to the service
  - **AWS Data Pipeline:** clusters scheduling and initiating
  - **Amazon VPC:** virtual network configurations for enhanced security.

- When configuring EMR, you can:
  - Use Spot instances and run task nodes
  - Use reserved instances with long running clusters

ELASTIC MAP REDUCE (EMR) – PART #2

WHAT IS SPOT INSTANCE?

- A Spot offers a lower price compared to an on-Demand instance.
- “Spot price” is the price you pay for the spot instance and is adjusted based on availably zone and demand.
- The spot instance will run when:
  - When capacity permits.
  - When the maximum hourly price set is more than the Spot price, the instance will run.
- Choose Spot instances when:
  - You have limited resources
  - When you have some degrees of flexibility. Note that your applications can be interrupted.
  - You want to perform optional tasks, data analysis, and batch jobs.

Photo Credit: https://pxhere.com/en/photo/747848

EMR DATA STORAGE

- Amazon EMR and Hadoop offer several file systems when processing a cluster (you can use multiple of them).
- HDFS and EMRFS are the most common with Amazon EMR.

File System	Prefix	Description
HDFS	hdfs://	• HDFS is a distributed, scalable, and portable file system for Hadoop. • Fast • Storage is reclaimed when cluster ends.
EMRFS (EMR FILE SYSTEM)	s3://	• EMRFS is an implementation of the Hadoop file system used for reading and writing regular files from Amazon EMR directly to Amazon S3. • EMRFS allows for storing files in S3 for use with Hadoop while allowing features like Amazon S3 server-side encryption. • EMRFS allows for EMRFS consistent view which is a feature that leverages DynamoDB to ensure data consistency.
local file system		• locally connected disk.
Amazon S3 block file system	s3bfs://	• legacy file storage system. • Not recommended.

EMRFS CONSISTENT VIEW

- EMRFS consistent view overcomes some of the issues encountered due to S3 Data Consistency model.
- Scenario: If data is added to Amazon S3 followed by another operation that include “list objects”, you many encounter issues such as incomplete list. More common in multistep ETL (Extract transform Load) operations.
- EMRFS consistent view overcomes this issue by allowing the EMR cluster to check for list and read after write consistency.
- EMRFS Consistent view ensures data consistency by using an Amazon DynamoDB database to store metadata and ensure consistency with S3.
- There is a fee associated with EMRFS consistent view.

EMR SECURITY

- EC2 firewall settings are automatically configured by EMR.
- Launches clusters Amazon Virtual Private Cloud (VPC)
- Server-side and client side encryption in S3 used with EMRFS (an object store for Hadoop on S3).
- AWS Key Management Service and IAM polices/roles.
- Allows for In-transit and at-rest encryption
- Authentication with Kerberos

EMR NOTEBOOKS

- Apache allows for using Amazon EMR Notebooks which are serverless Jupyter notebooks.
- EMR notebooks are launched via AWS EMR console.
- EMR notebooks are used with Amazon EMR clusters running Apache Spark.
- EMR notebooks contents such as equations, models, and code are saved on Amazon S3 separately from the cluster that runs the code.
- An Amazon EMR cluster is needed to run codes in an EMR notebook, but notebooks are not locked with a specific cluster.
- EMR Notebooks allow for cluster provisioning and could be hosted inside a VPC.
- Similar to Zepplin.

EMR SPOT INSTANCES

- Spot instances could be used for task nodes to save money.
- Spot instances are not recommended for core and master since it might result in data loss.

Master node:

- m4.large or m4.xlarge

Core & task nodes:

- m4.large is recommended

WHEN SHOULD I USE AWS GLUE VS. AWS EMR?

AWS GLUE:

- Glue is an ETL service that runs on a serverless Apache Spark environment.
- As a user, you do not have to configure or manage resources.
- Glue contains a data catalogue for ETL that could be used with Athena and Redshift Spectrum.
- AWS Glue ETL jobs uses Scala or Python.

AWS EMR:

- Amazon EMR allows for direct access to Hadoop environment.
- Amazon EMR allows for flexible, lower-level access to tools beyond Spark.

WHEN TO USE ATHENA COMPARED TO AMAZON EMR AND REDSHIFT?

Redshift:

- Data warehousing services
- Offers fastest query performance for enterprise reporting and business intelligence workloads
- Can be used with extremely complex SQL with several sub-queries

EMR:

- Sophisticated data processing framework
- Allows for running highly distributed processing frameworks such as Hadoop, Spark, and Presto in a simple and cost effective ways.
- Offers large flexibility, users are able to specify memory, compute and storage requirements to meet their specific needs.

Athena:

- Query service
- Provides easiest way to run ad-hoc serverless queries for data available in S3 buckets.
- There is absolutely no need to manage any servers.

APACHE HADOOP ON AMAZON EMR

APACHE HADOOP ON AMAZON EMR

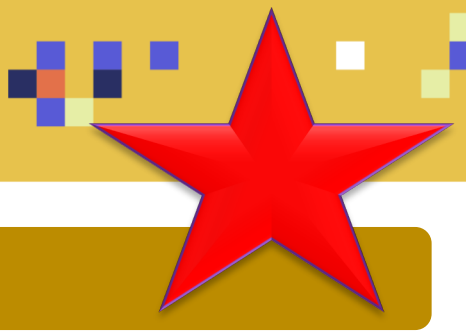
- Apache Hadoop is an open source tool to process large data efficiently.
- Hadoop allows for hardware clustering to process many datasets in parallel (instead of relying on one computer).
- Amazon EMR allows for launching elastic clusters of Amazon EC2 instances running Hadoop.
- Hadoop includes:
  - **MapReduce:** execution framework
  - **YARN:** resource manager
  - **HDFS:** distributed storage
- Amazon EMR includes EMRFS which allows Hadoop to use Amazon S3 for storage.
- Amazon EMR could be used to install tools such as Hive, Pig, Hue, Ganglia, Oozie, and HBase on your cluster.

APACHE HADOOP ON AMAZON EMR: HADOOP AND BIG DATA

- Hadoop is extremely scalable and process data in parallel which makes it perfectly suited for big data processing.
- Hadoop is very durable and available.
- Scalability is achieved by adding more servers to the Hadoop cluster.
- Amazon EMR could be used to create and configure a cluster of Amazon EC2 instances that run Hadoop.

Photo Credit: https://www.needpix.com/photo/514304/big-data-data-analysis-information-data-analysis-analytics-data-analytics

APACHE HADOOP ON AMAZON EMR: ELEMENTS



MAPREDUCE AND YARN

- Hadoop MapReduce execution engine works by dividing a job into smaller ones and distribute them across many nodes in Amazon EMR Cluster.
- YARN is a resource manage that tracks resources in a cluster, and it ensures that these resources are dynamically allocated.

STORAGE USING AMAZON S3 AND EMRFS

- EMR File System (EMRFS) could be used to leverage S3 storage for Hadoop. S3 is scalable and decoupled from compute resources.
- EMRFS is optimized for Hadoop to read and write in parallel to Amazon S3.
- Allows for object encryption with Amazon S3 for both server-side and client-side.

ON-CLUSTER STORAGE WITH HDFS

- Hadoop offers a distributed storage system named HDFS (the Hadoop Distributed File System) that stores data on disk on clusters in large blocks.
- HDFS offers a 3x replication factor for improved speed and availability.

APACHE HADOOP ON AMAZON EMR: UNIQUE FEATURES

HIGH SETUP SPEED

- Hadoop clusters could be launched in minutes using Amazon EMR cluster

MINIMUM ADMIN WORK

- Amazon EMR manages all the hassle associated with configuring and ensuring security of Hadoop.

SEAMLESS INTERGRATION WITH OTHER AWS SERVICES

- Hadoop is integrated with Amazon S3, Kinesis, Redshift, and DynamoDB.
- AWS Glue Data Catalog is used as a metadata repository for Apache Hive and Apache Spark.

COST EFFECTIVE

- Amazon EMR allows for launching dynamic instances (with auto scaling) to address varying workloads as opposed to the classic expensive capacity planning before deploying a Hadoop environment.

HIGH AVAILABILITY

- Hadoop on Amazon EMR could be launched in many availability zones and therefore ensures high availability.

APACHE HADOOP ON AMAZON EMR: USE CASES

CLICKSTREAMS DATA ANALYTICS

- Hadoop is used for clickstreams data analytics to track customers behaviour (targeted ad campaign).

LOG DATA ANALYSIS

- Hadoop is used for log data analysis by converting petabytes of un-structured data into key metrics/insights.

MASSIVE DATA PROCESSING

- Hive application can allow for massive scale data analytics by leveraging MapReduce using a SQL interface.

ETL JOBS

- Hadoop works great for ETL jobs such as sorting, joining, and aggregating big data.

APACHE SPARK ON AMAZON EMR

APACHE SPARK

- Apache Spark is an open-source, widely adopted big data processing system.
- It performs batch processing, optimization and in-memory caching for extreme performance and minimum latency.
- Spark is supported in Amazon EMR
- You can launch Apache Spark Clusters in minutes from the console besides leveraging the following features:
    - S3 storage and connectivity using Amazon EMR File System (EMRFS)
    - EC2 spot instances
    - AWS Glue to store Spark SQL table metadata
    - Auto scaling to account for dynamic demand
- Spark encryption and authentication is configured with Kerberos using an Amazon EMR security configuration.
- Amazon EMR installs and manages Apache Spark on Hadoop YARN.

APACHE SPARK: FEATURES

SUPER FAST

- Apache Spark can perform extremely fast data transformation.
- Apache Spark achieves fast processing by removing I/O cost. This is achieved by storing inputs and outputs in-memory as resilient distributed datasets (RDDs).

EASY TO USE

- Apache Spark works with many languages such as Java, Scala, and Python.
- SQL or HiveQL queries could be sent to Apache Spark via Spark SQL module.
- Zeppelin could be used to develop interactive notebooks to visualize your data (similar to jupyter notebooks).

DIVERSE WORKFLOWS

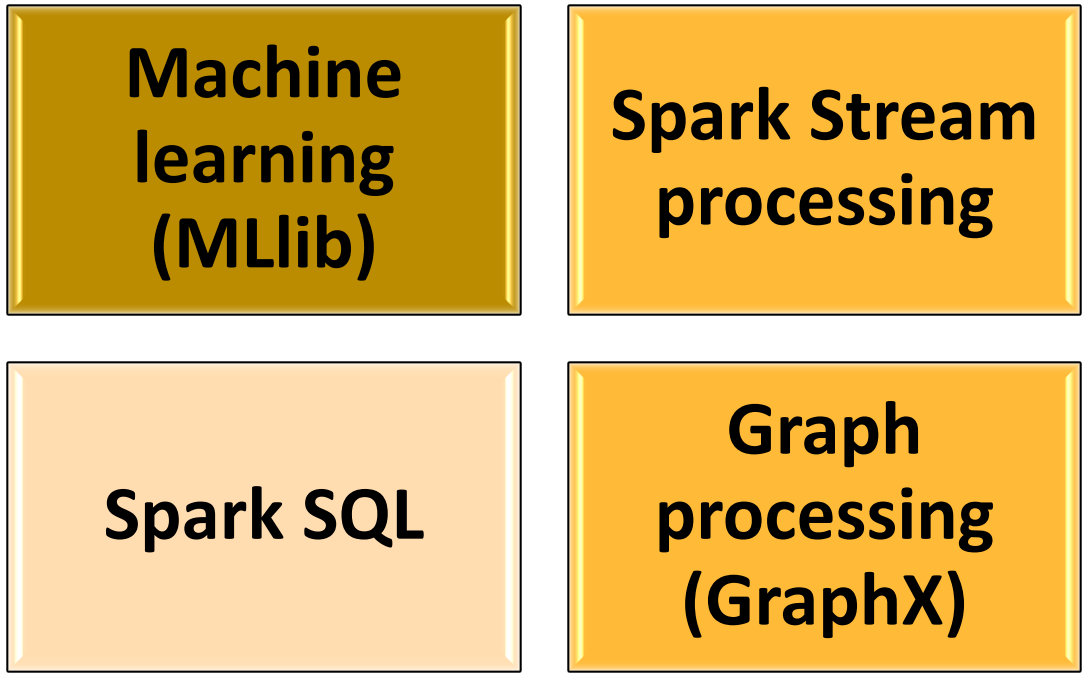
- Several libraries are included such as machine learning (MLlib), stream processing (Spark Streaming), and graph processing (GraphX).
- Apache MXNet could be used as well for Deep Learning applications.

EASY EMR INTEGRATION

- Submit Apache Spark jobs with the Amazon EMR Step API, use Apache Spark with EMRFS to directly access data in Amazon S3, save costs using Amazon EC2 Spot capacity.

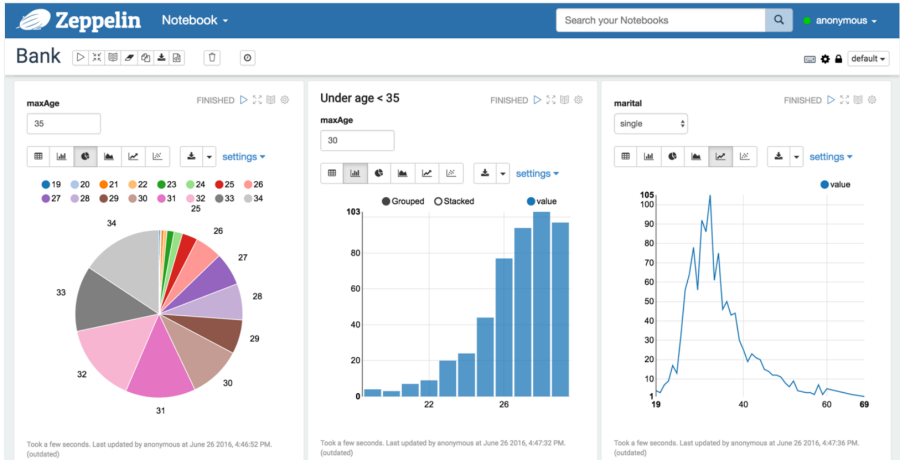
APACHE SPARK

- Apache Spark includes several libraries such as machine learning (MLlib), stream processing (Spark Streaming), and graph processing (GraphX).



APACHE ZEPPLIN

- Apache Zepplin is a web-based notebook that allows for interactive data analytics and collaborative documents with SQL, Scala.
- Zepplin is a Multi-purpose Notebook that could be used for:
  - Discovery
  - Ingestion
  - Analytics
  - Visualization
- Apache Zeppelin could be used for data exploration by creating interactive notebooks using Apache Spark.
- Zepplin could be used as well with deep learning frameworks like Apache MXNet with Spark applications.



APACHE SPARK: USE CASES

DATA STREAMING

- Apache spark could be integrated with Amazon Kinesis and Apache Kafka to stream and analyze data in Realtime.
- Results could be stored to Amazon S3 or on-cluster HDFS.

MACHINE LEARNING

- MLlib library is available with Apache Spark on Amazon EMR to train Machine learning models

SQL

- Spark SQL could send queries with SQL or HiveQL with very low-latency.
- Apache Spark on Amazon EMR can leverage EMRFS to access data on Amazon S3.
- Zeppelin notebooks.
- BI tools via ODBC and JDBC connections.