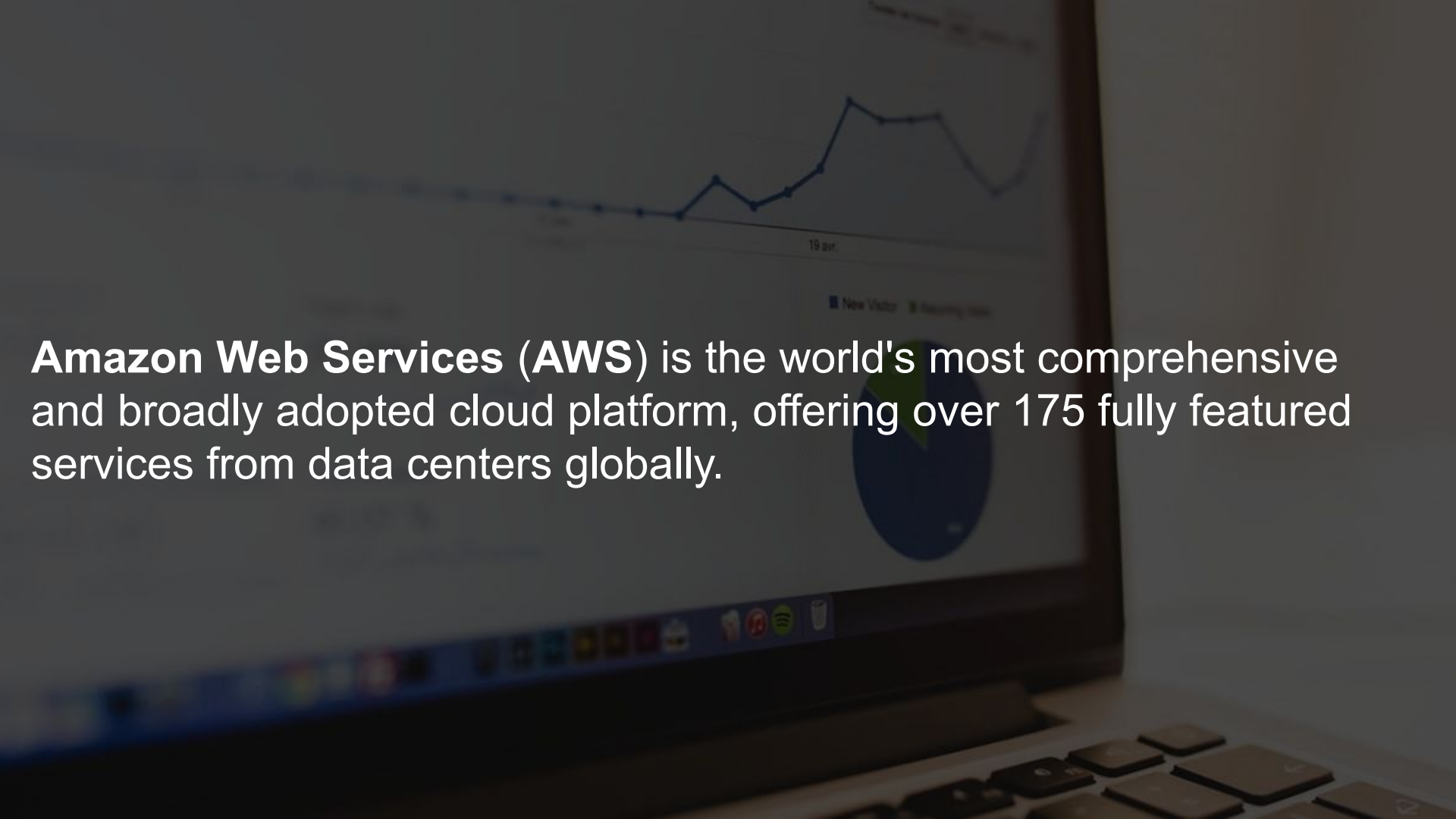
Cloud is about how you do computing, not where you do computing.



Amazon Web Services (**AWS**) is the world's most comprehensive and broadly adopted cloud platform, offering over 175 fully featured services from data centers globally.

What AWS do?

Amazon Web Services offers a broad set of global cloud-based products including compute, **storage**, databases, analytics, networking, mobile, developer tools, management tools, IoT, security and enterprise applications.

These services help organizations move faster, lower IT costs, and scale.

Big Data Services

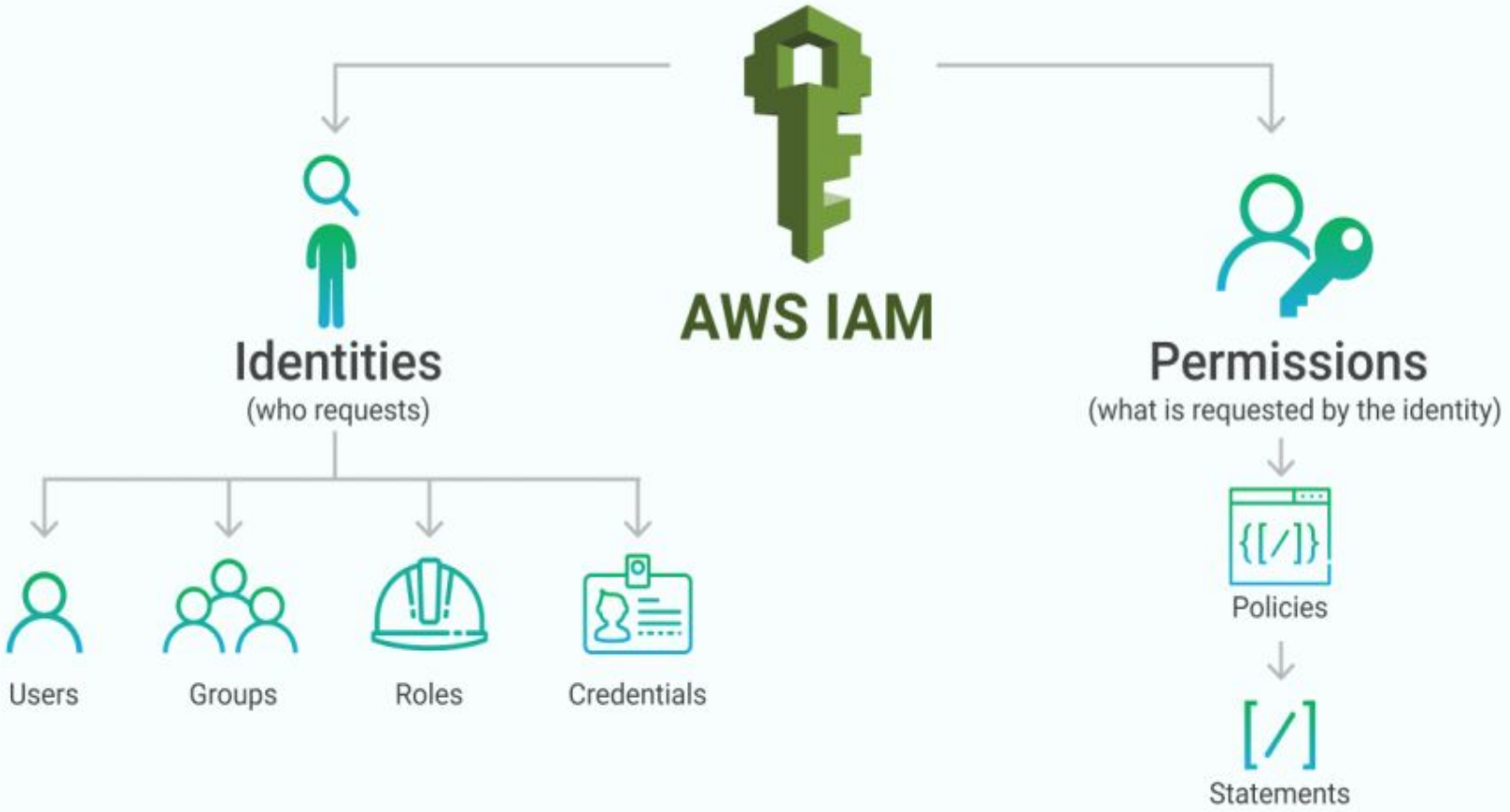
- AWS IAM
- AWS S3
- AWS Redshift
- AWS EC2
- AWS Glue
- AWS Kinesis
- AWS EMR (Spark)
- AWS Lambda

AWS IAM

AWS Identity and Access Management (IAM) enables you to manage access to AWS services and resources securely. Using IAM, you can create and manage AWS users and groups, and use permissions to allow and deny their access to AWS resources.

IAM is a feature of your AWS account offered at no additional charge. You will be charged only for use of other AWS services by your users.

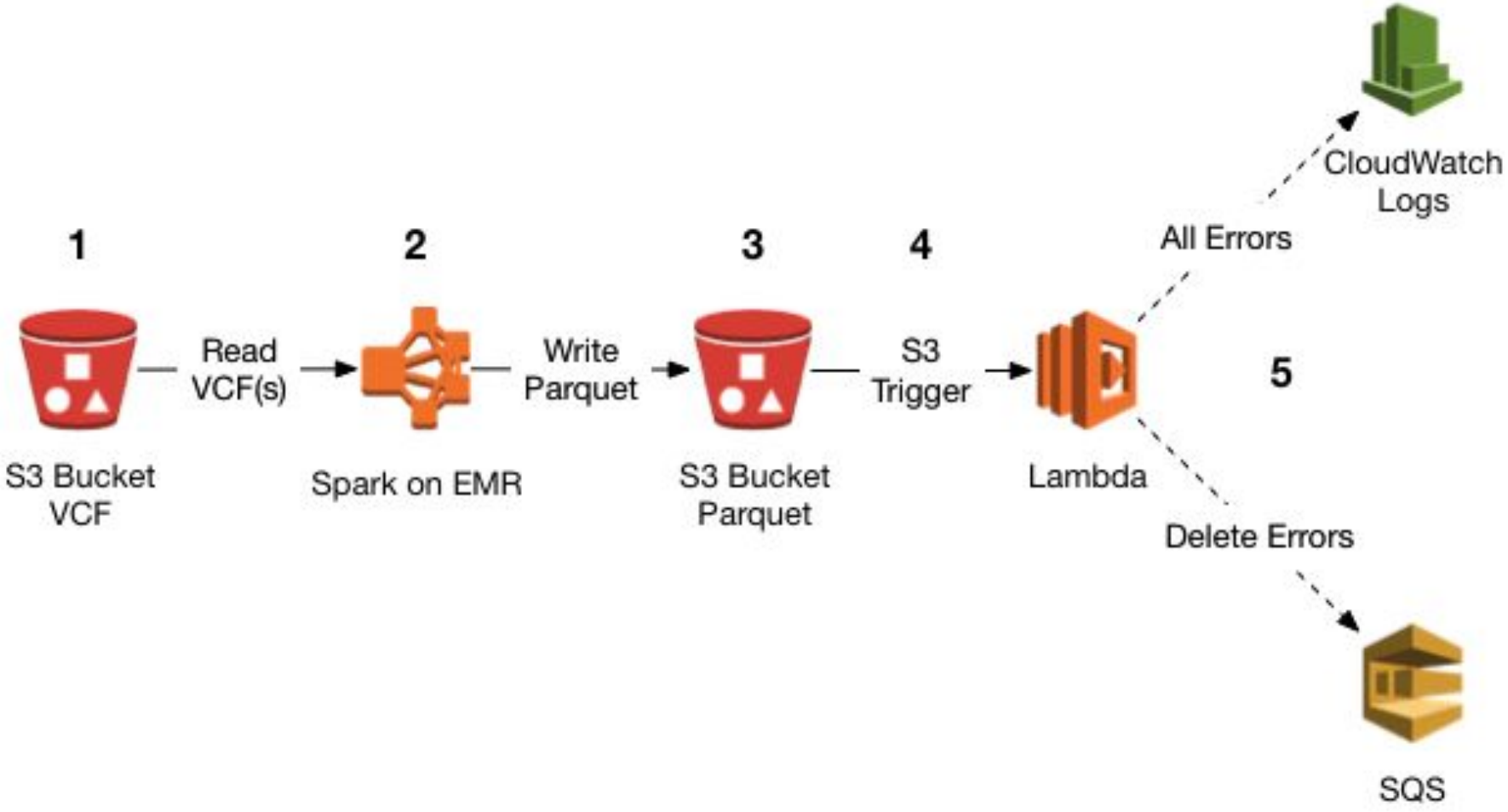
- IAM Group
- IAM Users
- IAM Roles


AWS IAM
Identities
(who requests)
Users
Groups
Roles
Credentials
Permissions
(what is requested by the identity)
Policies
Statements

AWS S3

Amazon Simple Storage Service (Amazon S3) is an object storage service that offers industry-leading scalability, data availability, security, and performance. This means customers of all sizes and industries can use it to store and protect any amount of data for a range of use cases, such as data lakes, websites, mobile applications, backup and restore, archive, enterprise applications, IoT devices, and big data analytics.

Amazon S3 provides easy-to-use management features so you can organize your data and configure finely-tuned access controls to meet your specific business, organizational, and compliance requirements. Amazon S3 is designed for 99.999999999% (11 9's) of durability, and stores data for millions of applications for companies all around the world.


1
2
3
4
5
S3 Bucket
VCF
Read
VCF(s)
Spark on EMR
Write
Parquet
S3 Bucket
Parquet
S3
Trigger
Lambda
All Errors
CloudWatch
Logs
Delete Errors
SQS

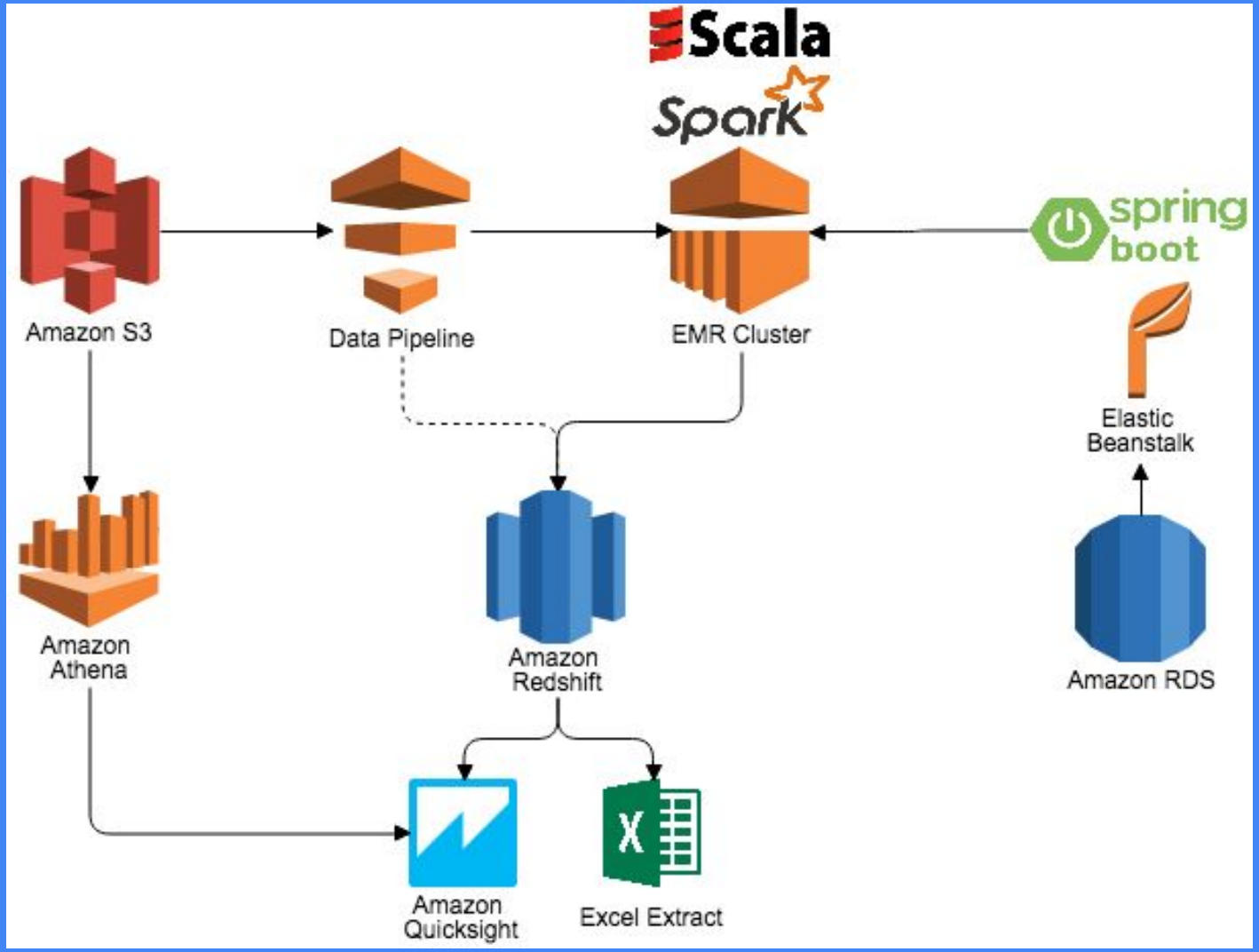
AWS Redshift

The most popular and fastest cloud data warehouse.

Redshift powers analytical workloads for Fortune 500 companies, startups, and everything in between. Companies like Lyft have grown with Redshift from startups to multi-billion dollar enterprises.

Amazon Redshift is a fully managed, petabyte-scale data warehouse service in the cloud.This enables you to use your data to acquire new insights for your business and customers. The first step to create a data warehouse is to launch a set of nodes, called an **Amazon Redshift** cluster.

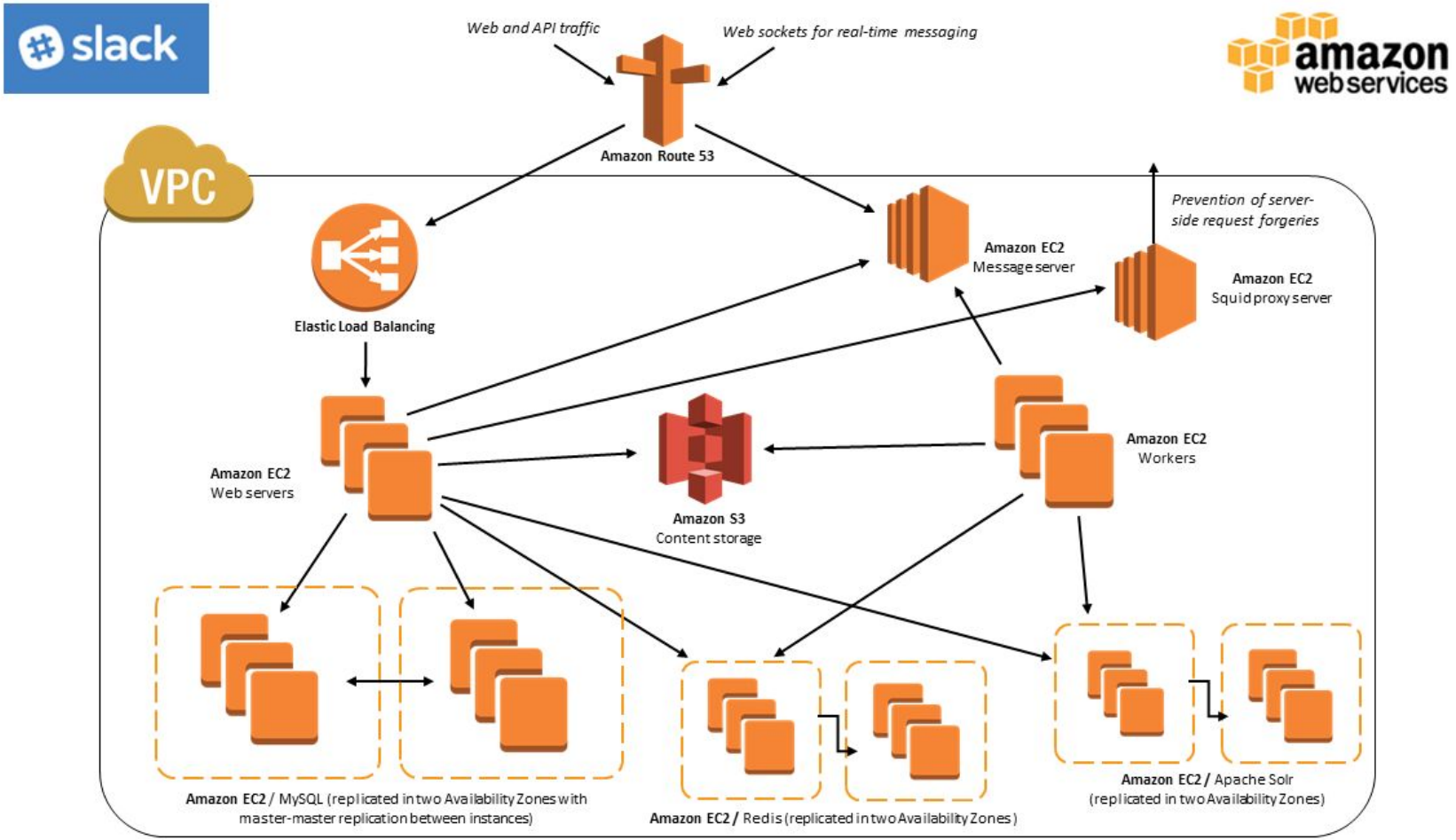
Amazon **Redshift** is **based on** industry-standard PostgreSQL, so most existing SQL client applications will work with only minimal changes.


Scala
Spark
spring boot
Amazon S3
Data Pipeline
EMR Cluster
Elastic Beanstalk
Amazon Athena
Amazon Redshift
Amazon RDS
Amazon Quicksight
Excel Extract

AWS EC2

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides secure, resizable compute capacity in the cloud. It is designed to make web-scale cloud computing easier for developers. Amazon EC2's simple web service interface allows you to obtain and configure capacity with minimal friction. It provides you with complete control of your computing resources and lets you run on Amazon's proven computing environment.

Amazon EC2 offers the broadest and deepest compute platform with choice of processor, storage, networking, operating system, and purchase model. We offer the fastest processors in the cloud and we are the only cloud with 400 Gbps ethernet networking. We have the most powerful GPU instances for machine learning training and graphics workloads, as well as the lowest cost-per-inference instances in the cloud. More SAP, HPC, Machine Learning, and Windows workloads run on AWS than any other cloud. Click here to learn What's New with Amazon EC2.



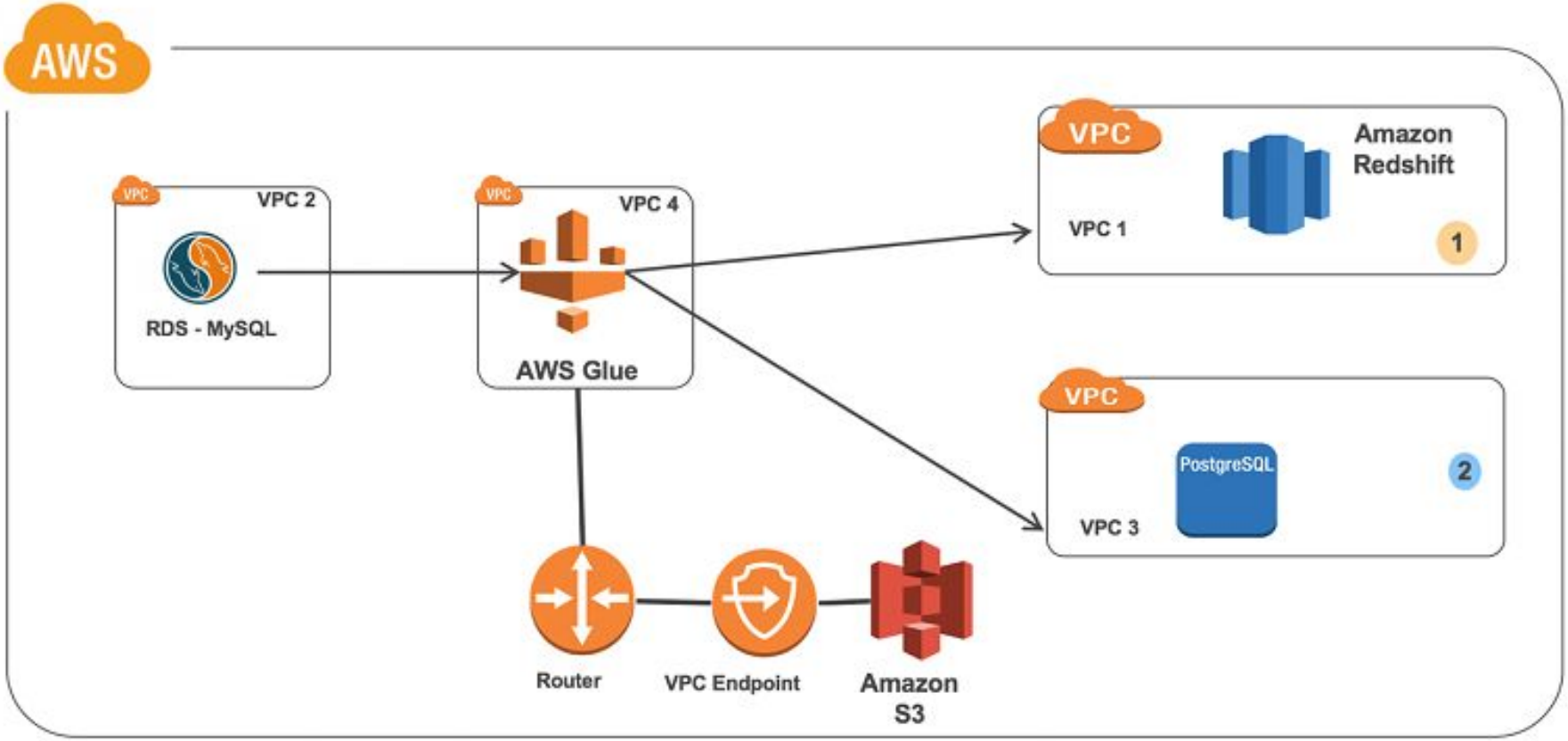
An overview of the Slack architecture on AWS

AWS Glue

AWS Glue is a serverless data preparation service that makes it easy for data engineers, extract, transform, and load (ETL) developers, data analysts, and data scientists to extract, clean, enrich, normalize, and load data. AWS Glue reduces the time it takes to start analyzing your data from months to minutes.

Data preparation is a critical but challenging process. To get data ready for analysis, you first extract data from various sources. You then clean it, transform it into the required format, and load it into databases, data warehouses, and data lakes for further analysis. These tasks are often performed by different groups with different tools.

AWS Glue provides you with both visual and code-based interfaces to make data preparation easy. Data engineers and ETL developers can use AWS Glue Studio to create, run, and monitor ETL workflows with a few clicks. Data analysts and data scientists can use AWS Glue DataBrew to visually clean up and normalize data without writing code.


AWS
VPC
VPC 2
RDS - MySQL
VPC
VPC 4
AWS Glue
VPC
Amazon
Redshift
VPC 1
1
VPC
PostgreSQL
2
VPC 3
Router
VPC Endpoint
Amazon
S3

AWS Kinesis

Amazon Kinesis makes it easy to collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information.

Amazon Kinesis offers key capabilities to cost-effectively process streaming data at any scale, along with the flexibility to choose the tools that best suit the requirements of your application.

With **Amazon Kinesis**, you can ingest real-time data such as video, audio, application logs, website clickstreams, and IoT telemetry data for machine learning, analytics, and other applications.

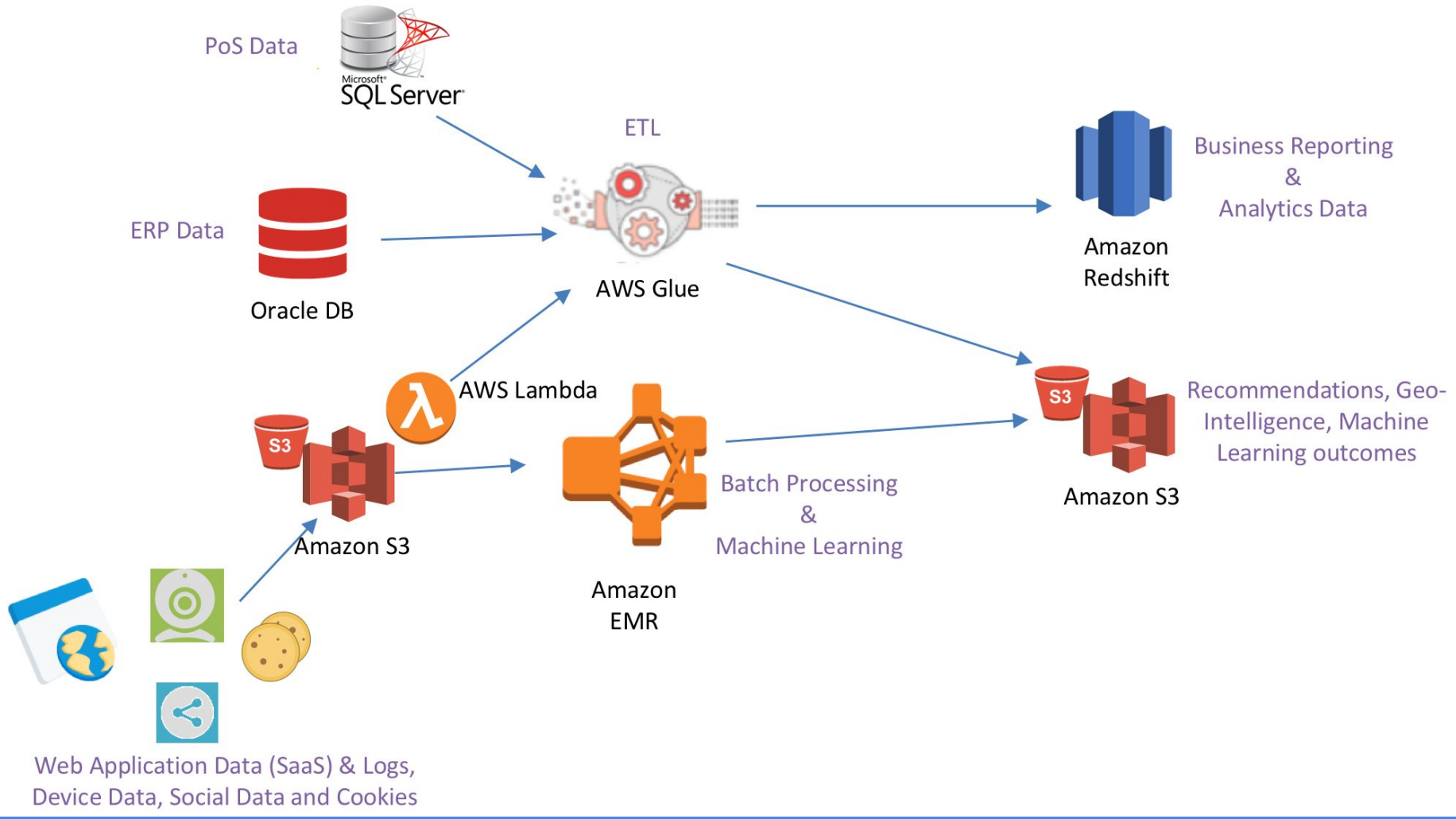
Amazon Kinesis enables you to process and analyze data as it arrives and respond instantly instead of having to wait until all your data is collected before the processing can begin.


Input
Websites send clickstream data to Amazon Kinesis Data Firehose
Amazon Kinesis Data Firehose
Collects the data and sends it to Amazon Kinesis Data Analytics
Amazon Kinesis Data Analytics
Processes data in real-time
Amazon Kinesis Data Firehose
Loads processed data into Amazon Redshift
Amazon Redshift
Run analytics models that come up with content recommendations
Output
Readers see personalized content suggestions and engage more

AWS EMR

Amazon EMR is the industry-leading cloud big data platform for processing vast amounts of data using open source tools such as Apache Spark, Apache Hive, Apache HBase, Apache Flink, Apache Hudi, and Presto.

With **EMR** you can run Petabyte-scale analysis at less than half of the cost of traditional on-premises solutions and over 3x faster than standard **Apache Spark**. For short-running jobs, you can spin up and spin down clusters and pay per second for the instances used. For long-running workloads, you can create highly available clusters that automatically scale to meet demand. If you have existing on-premises deployments of open source tools such as Apache Spark and Apache Hive, you can also run EMR clusters on AWS Outposts.


PoS Data
Microsoft SQL Server
ETL
ERP Data
Oracle DB
AWS Glue
Business Reporting
&
Analytics Data
Amazon
Redshift
AWS Lambda
S3
Amazon S3
Batch Processing
&
Machine Learning
Amazon
EMR
S3
Amazon S3
Recommendations, Geo-
Intelligence, Machine
Learning outcomes
Web Application Data (SaaS) & Logs,
Device Data, Social Data and Cookies

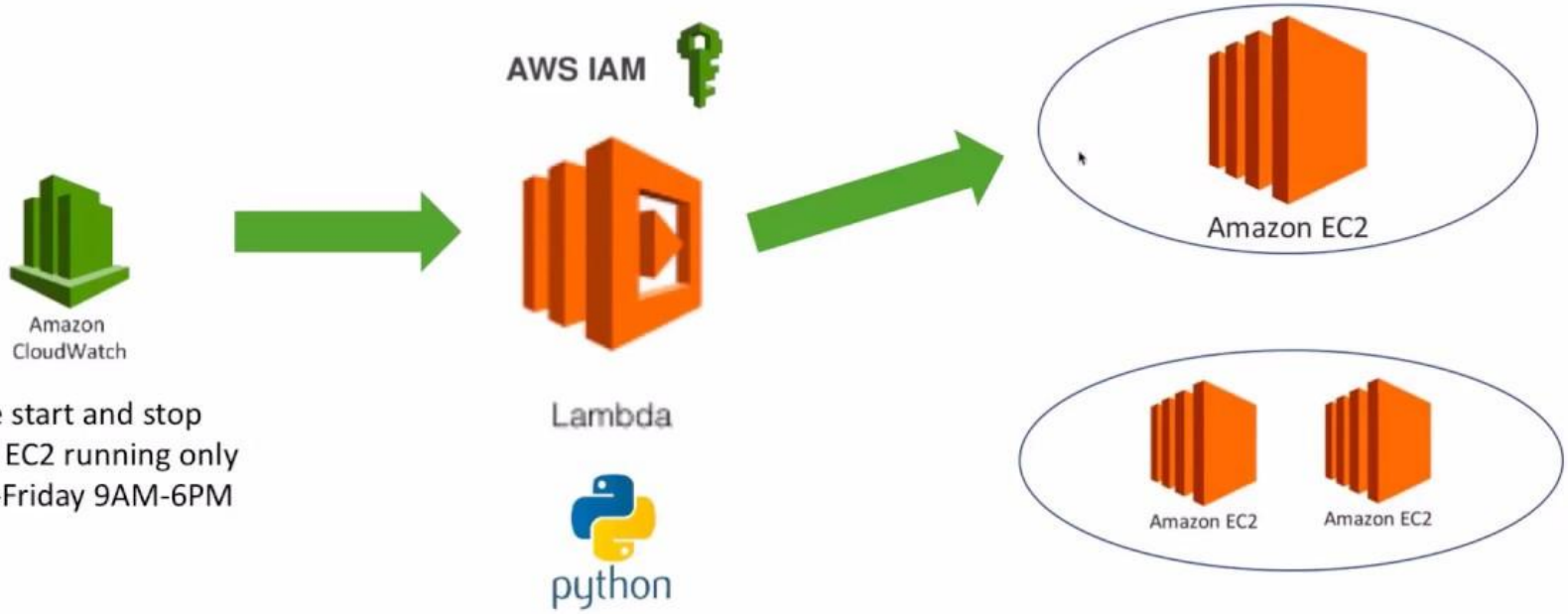
AWS Lambda

AWS Lambda lets you run code without provisioning or managing servers. You pay only for the compute time you consume.

With Lambda, you can run code for virtually any type of application or backend service, all with zero administration.

Just upload your code and **Lambda** takes care of everything required to run and scale your code with high availability. You can set up your code to automatically trigger from other AWS services or call it directly from any web or mobile app.

Start & Stop EC2 instances with AWS Lambda



Schedule start and stop
Eg: Keep EC2 running only
Monday-Friday 9AM-6PM


aws