Serving Business Insights with the AWS Analytics Suite (Level 200)

Ekta Parashar, Solution Architect Manager

What to Expect from this Session

- AWS Toolkit for advanced analytics
- Understand Stakeholders
- Services for Stakeholders
- Q & A

Who Am I ?

- Solution Architect at AWS since 2015.
- Based in Mumbai, India.

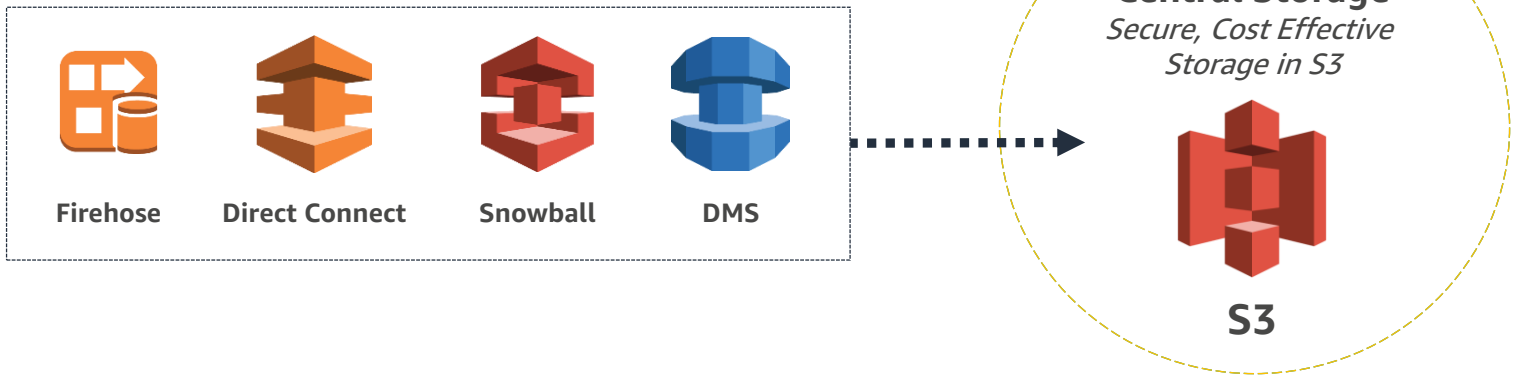
Central Storage

Secure, Cost Effective Storage in S3

S3

Data Ingestion

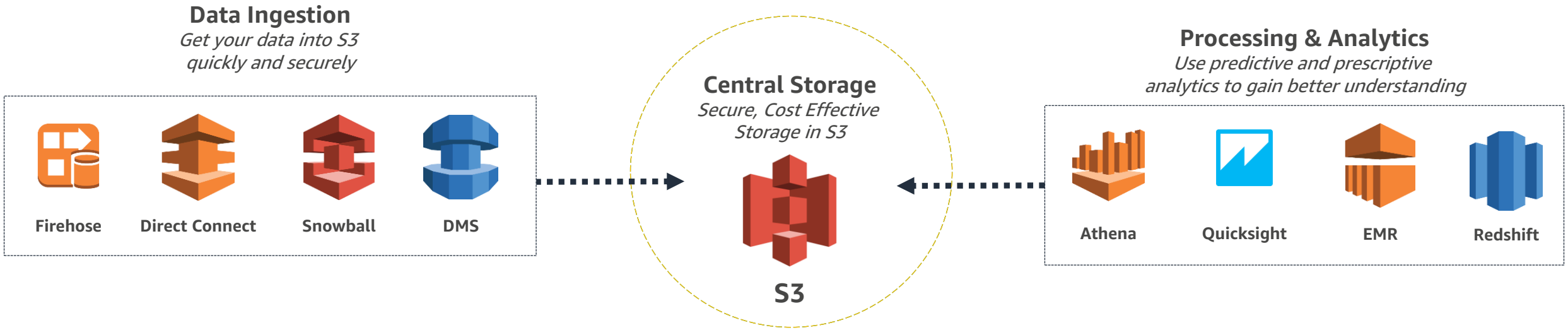
Get your data into S3 quickly and securely

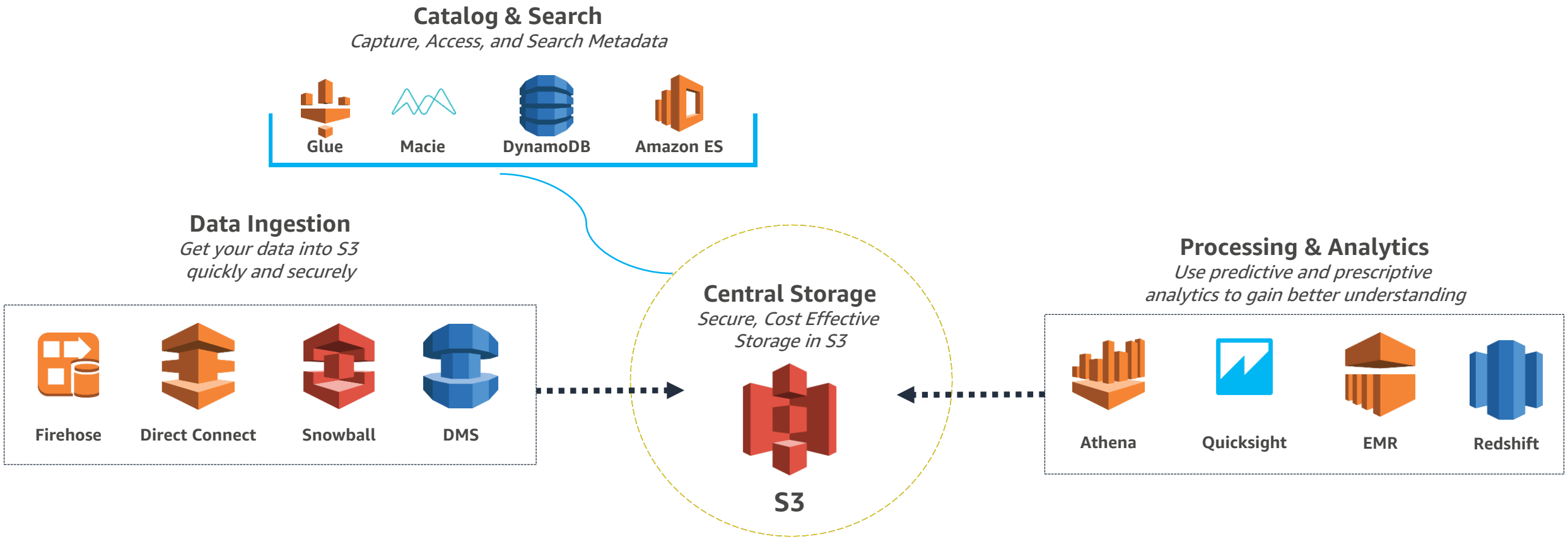


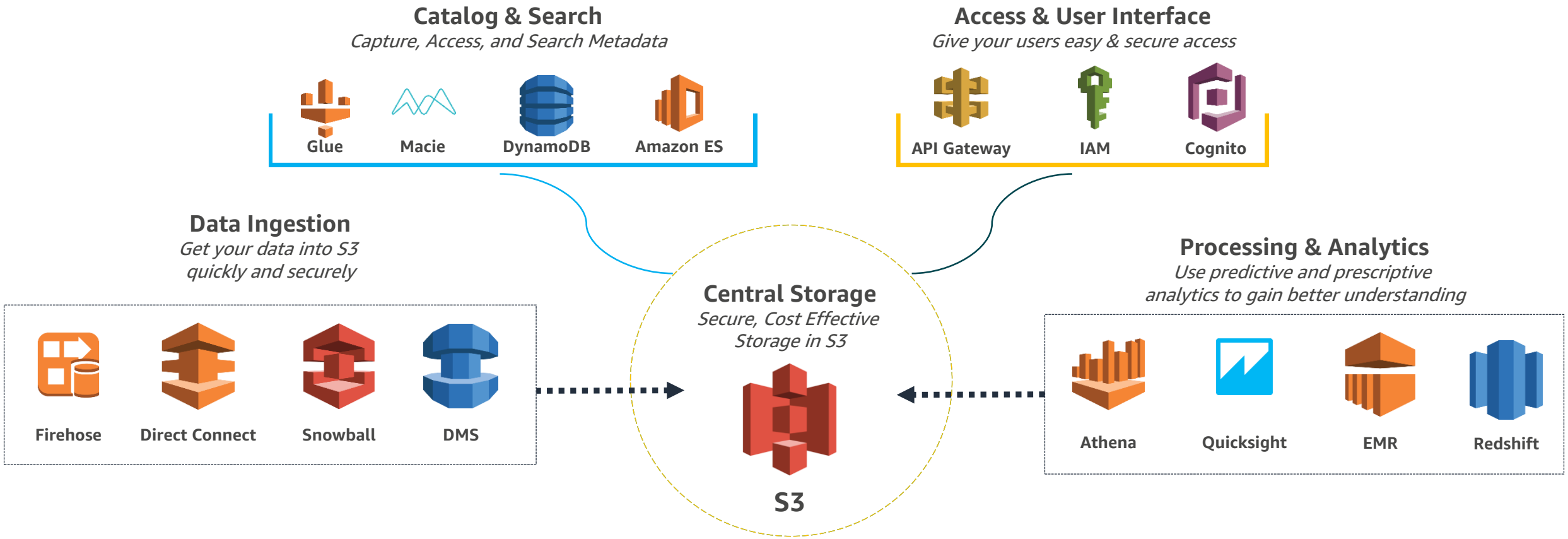
Central Storage

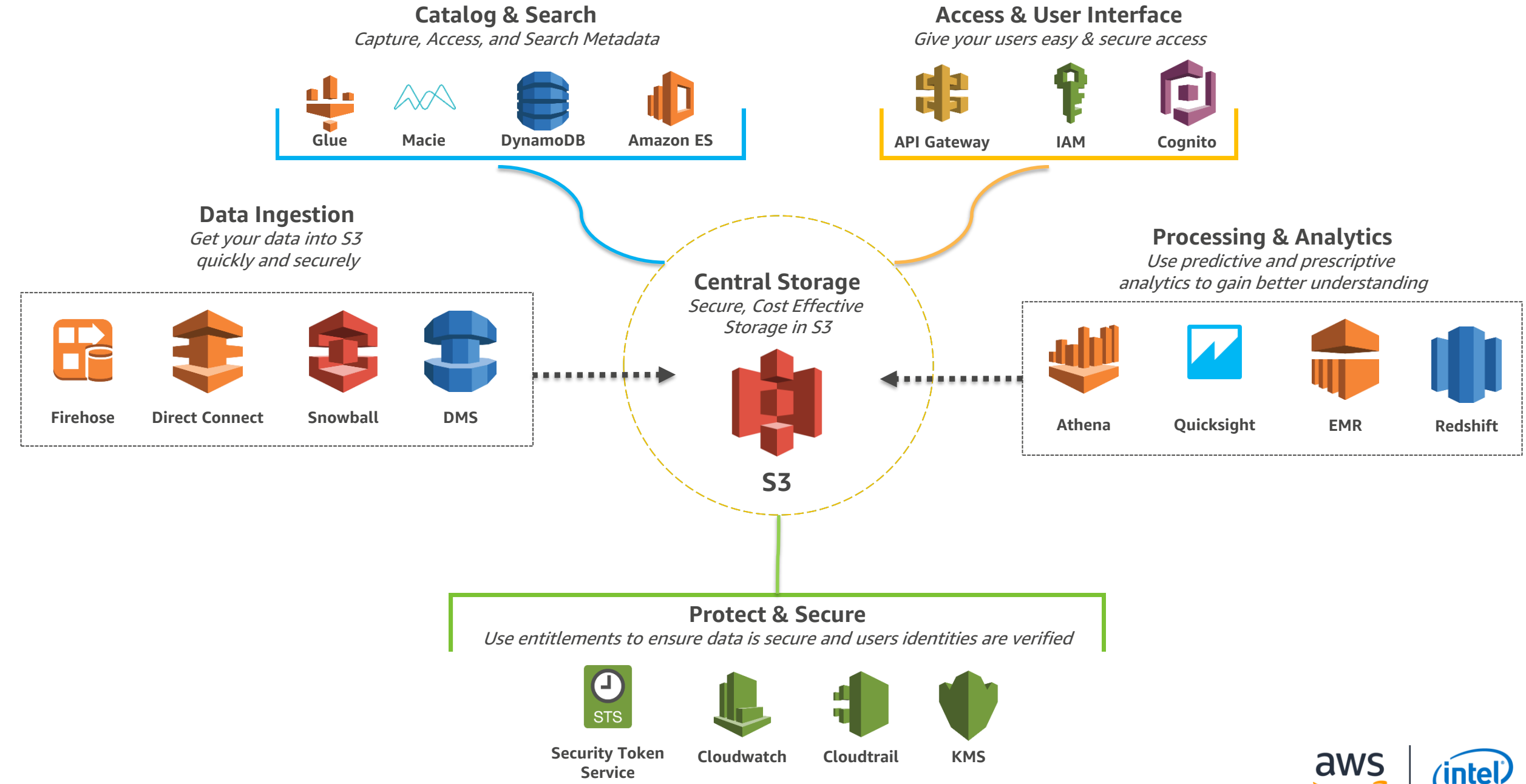
Secure, Cost Effective Storage in S3

S3

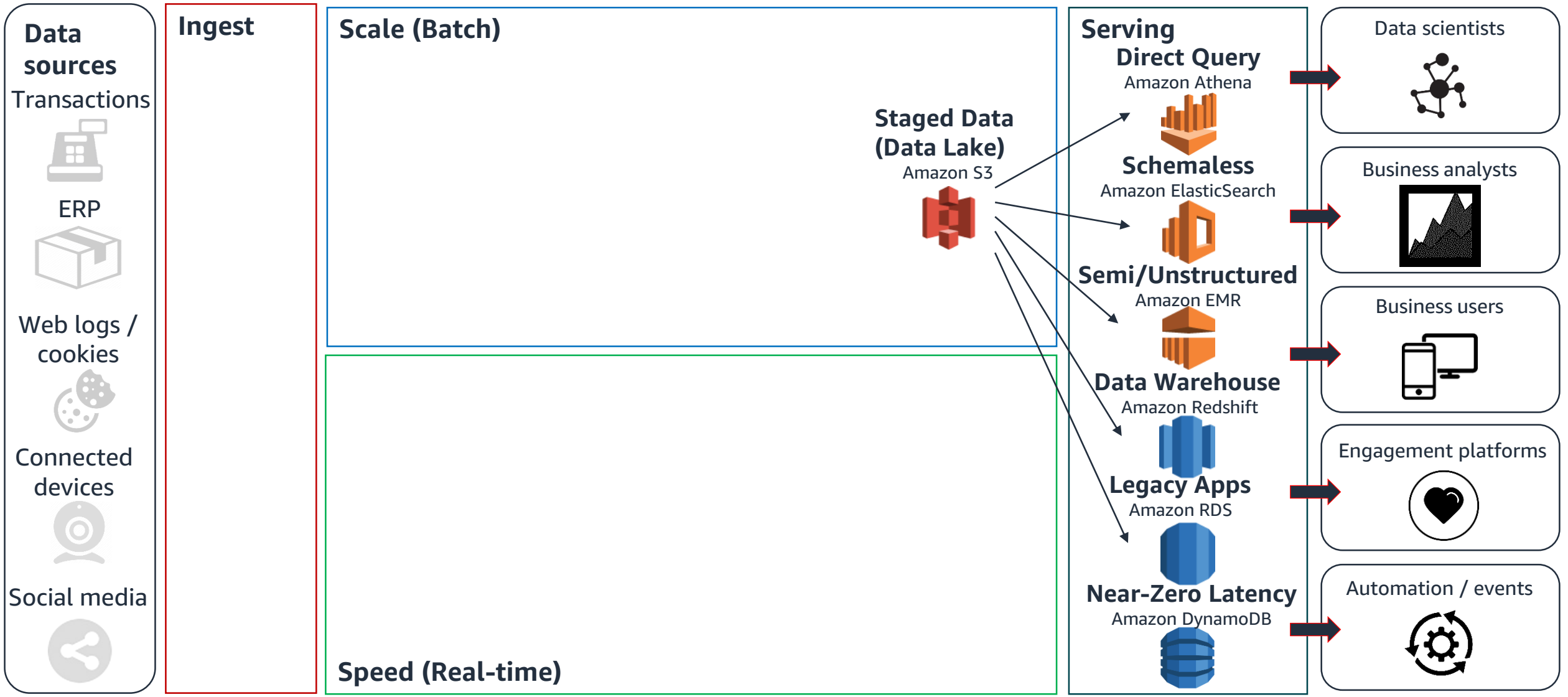








Modern data architecture
Insights to enhance business applications, new digital services

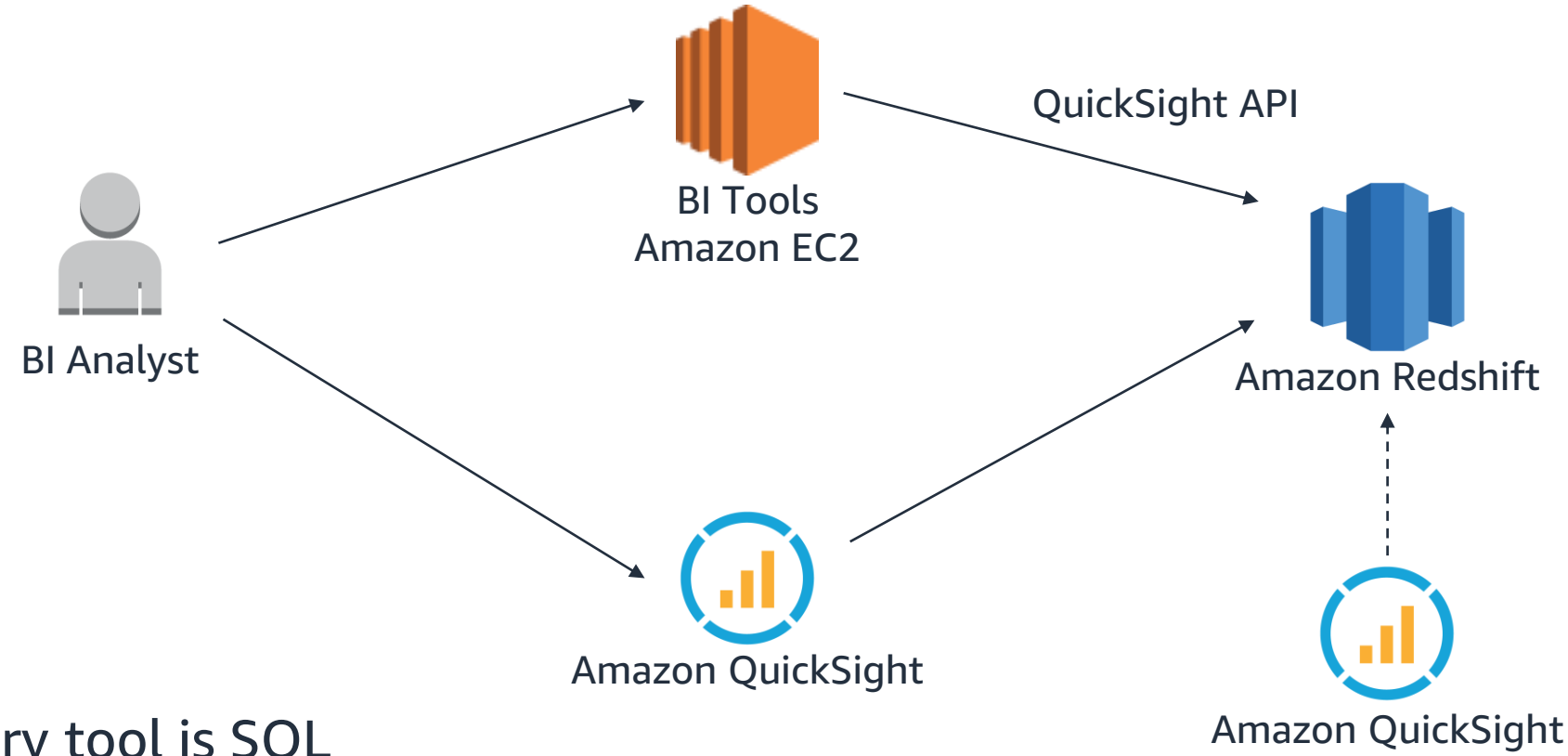


Match toolset to right persona

- **Business Intelligence (BI) Analyst**
  - Primary tool is SQL
  - Historical data resides in data warehouse such as Redshift, Spectrum, QuickSight
- **Data Scientist** – Uses programmatic languages such as R or python, explore data using Athena
- **Data Engineer**
  - Familiar with Hadoop and Spark
  - Wants to access data in Redshift in diverse ways

BI Analyst

BI Analyst with existing BI Tools



- Primary tool is SQL
- Data is largely structured with well known data sources
- Primary concern is fast, consistent performance
- Need to extend SQL with custom functions

Amazon Redshift system architecture

Leader node

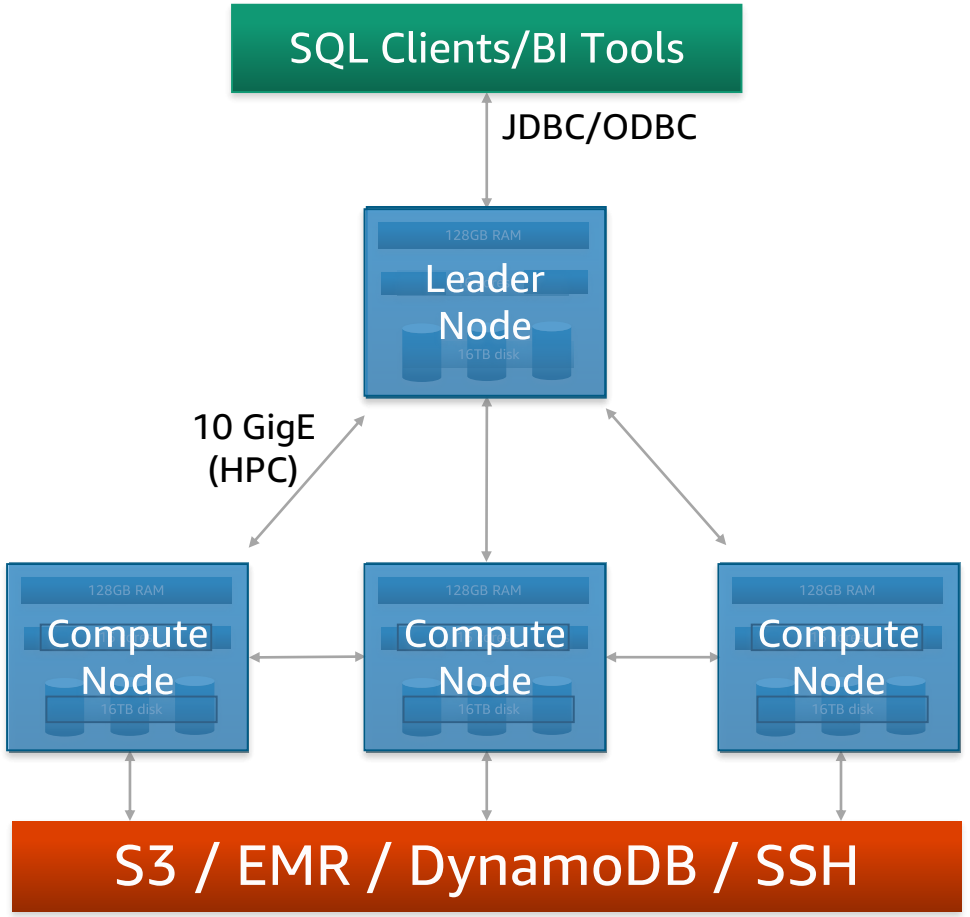
- SQL endpoint
- Stores metadata
- Coordinates query execution

Compute nodes

- Local, columnar storage
- Execute queries in parallel
- Load, backup, restore via Amazon S3; load from Amazon DynamoDB, Amazon EMR, or SSH

Two hardware platforms

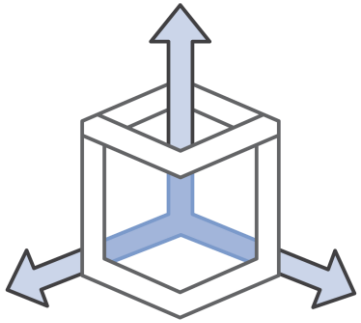
- Optimized for data processing
- DS2: HDD; scale from 2TB to 2PB
- DC1: SSD; scale from 160GB to 356TB



Amazon Redshift Spectrum

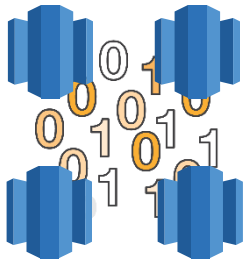
Run SQL queries directly against data in S3 using thousands of nodes

Fast @ Exabyte scale



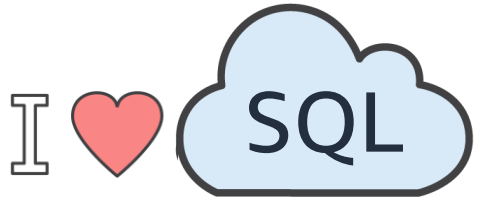
Elastic & highly available

On-demand, pay-per-query



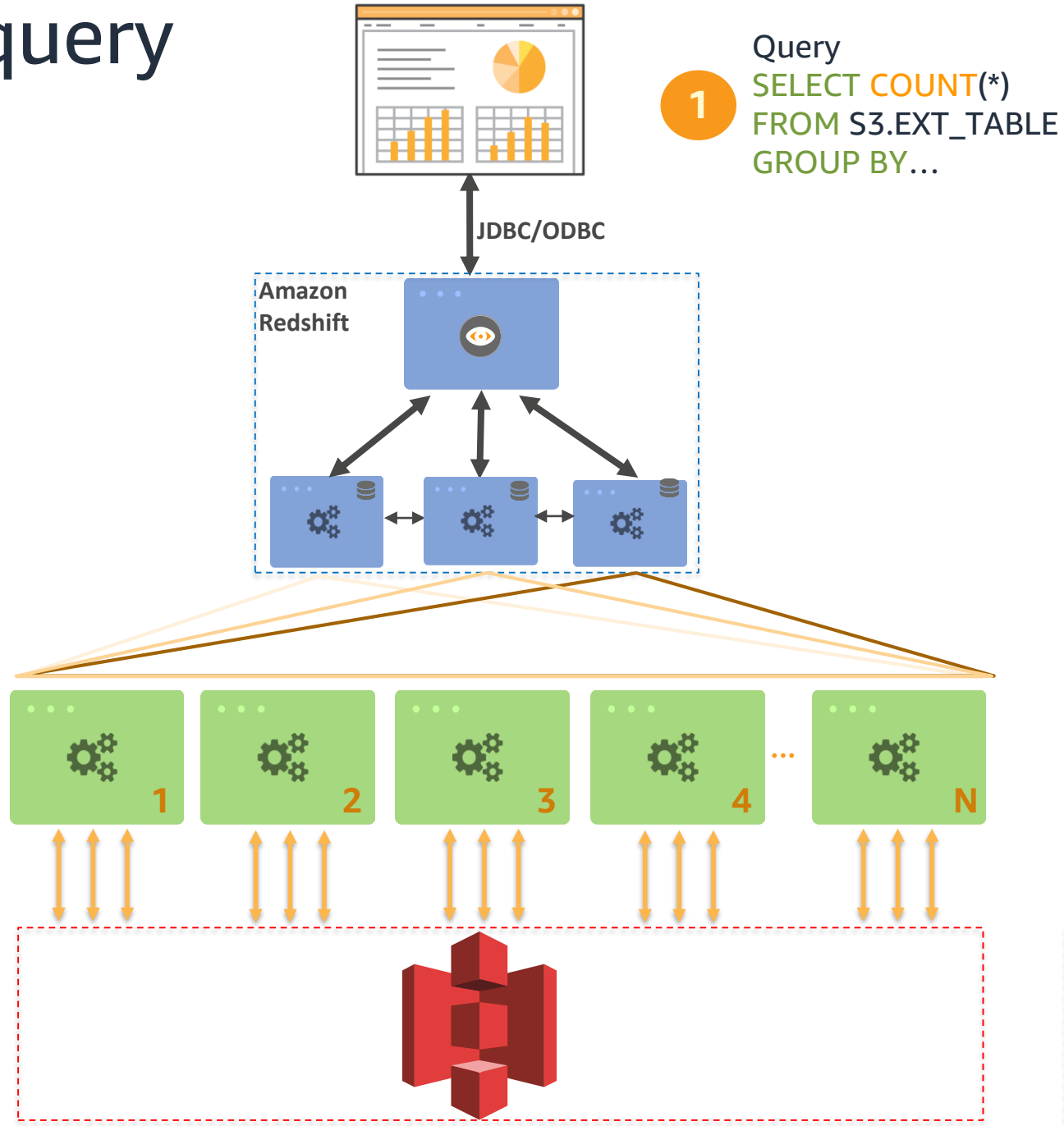
High concurrency: Multiple clusters access same data

No ETL: Query data in-place using open file formats



Full Amazon Redshift SQL support

Life of a query



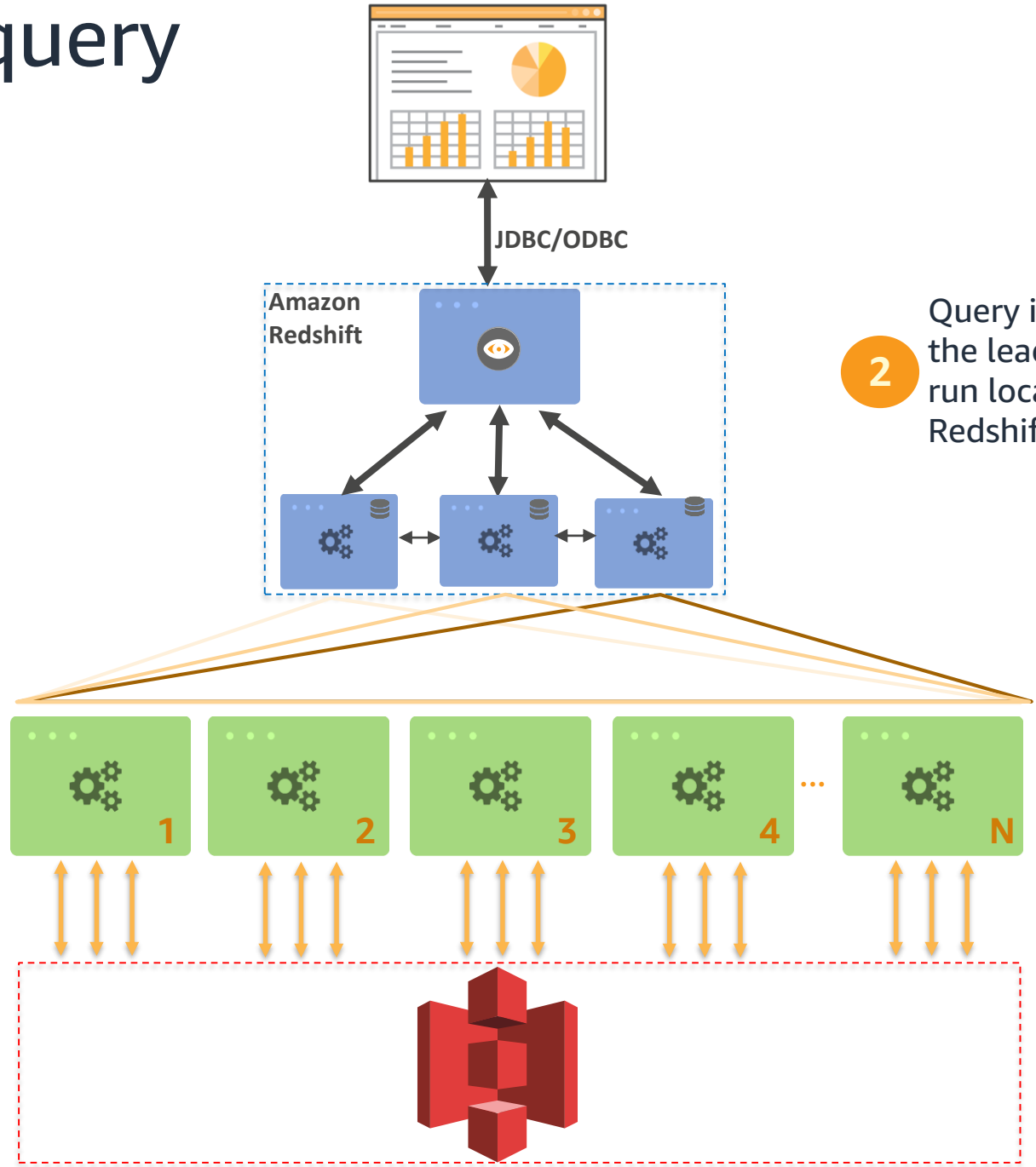
Amazon S3

Exabyte-scale object storage

Data Catalog

Apache Hive Metastore

Life of a query



2 Query is optimized and compiled at the leader node. Determine what gets run locally and what goes to Amazon Redshift Spectrum

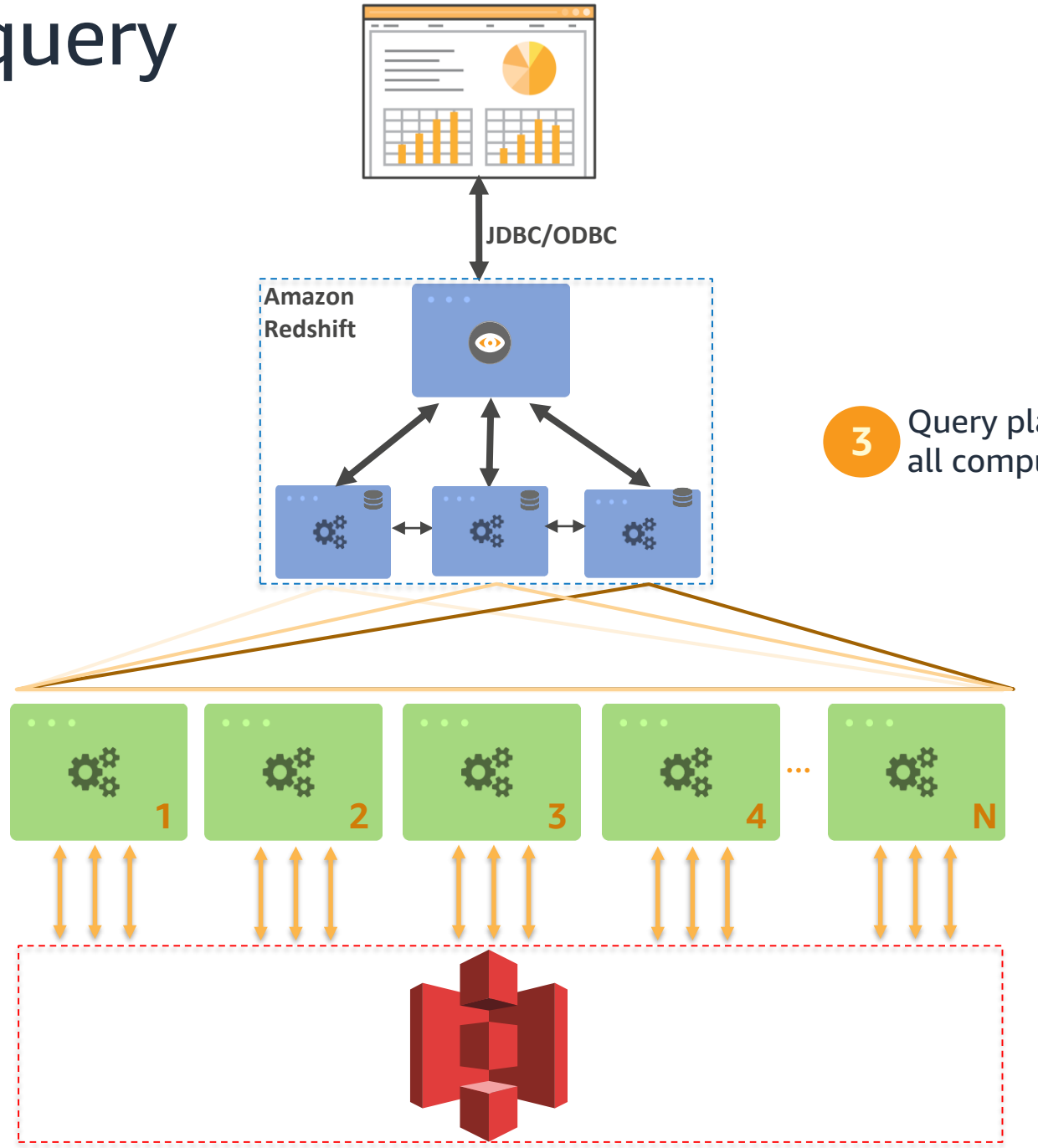
Amazon S3

Exabyte-scale object storage

Data Catalog

Apache Hive Metastore

Life of a query



3 Query plan is sent to all compute nodes

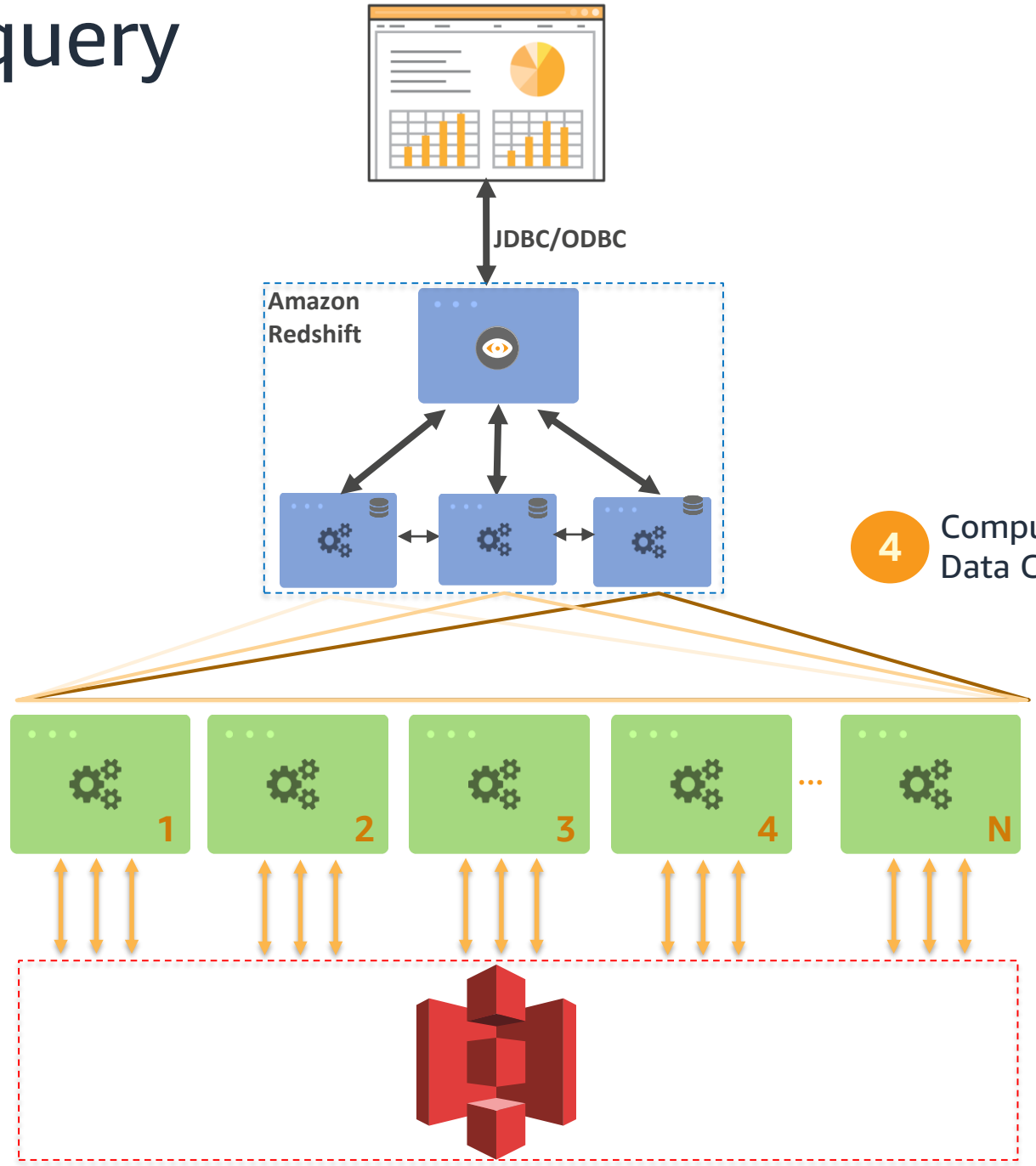
Amazon S3

Exabyte-scale object storage

Data Catalog

Apache Hive Metastore

Life of a query



4 Compute nodes obtain partition info from Data Catalog; dynamically prune partitions

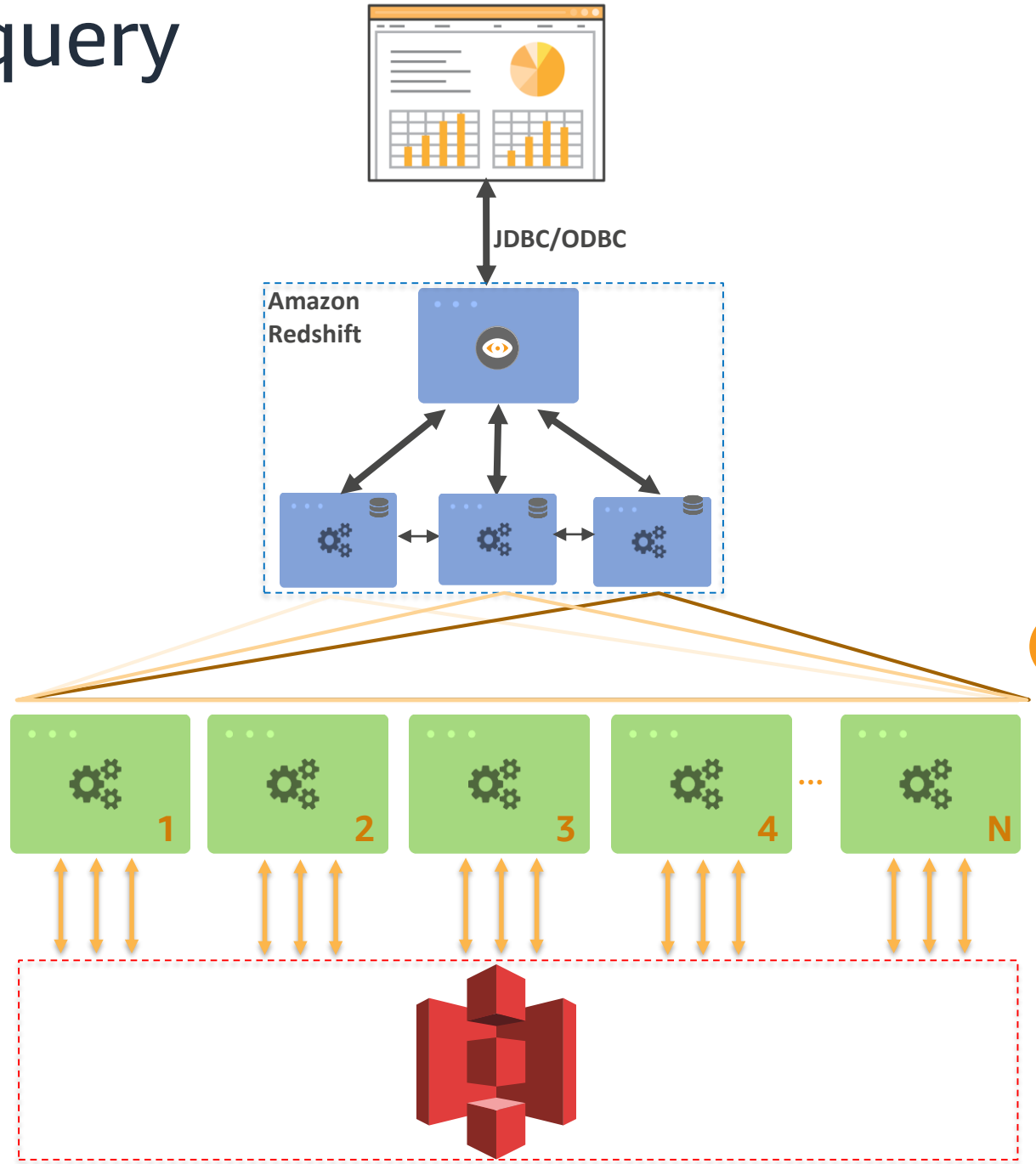
Amazon S3

Exabyte-scale object storage

Data Catalog

Apache Hive Metastore

Life of a query



5 Each compute node issues multiple requests to the Amazon Redshift Spectrum layer

Amazon S3

Exabyte-scale object storage

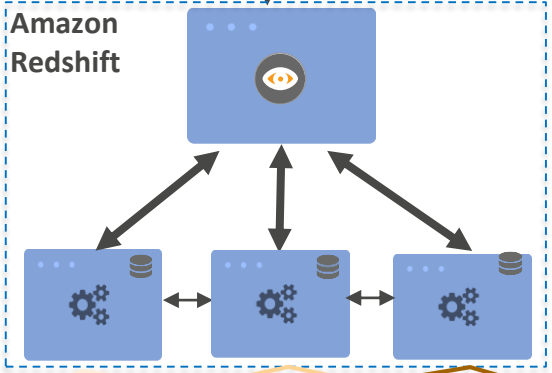
Data Catalog

Apache Hive Metastore

Life of a query

JDBC/ODBC

Amazon Redshift

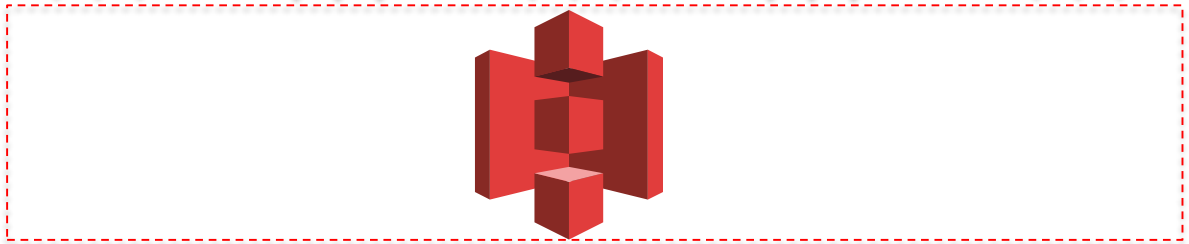


9 Result is sent back to client

8 Final aggregations and joins with local Amazon Redshift tables done in-cluster

7 Amazon Redshift Spectrum projects, filters, and aggregates

6 Amazon Redshift Spectrum nodes scan your S3 data



Amazon S3

Exabyte-scale object storage

Data Catalog

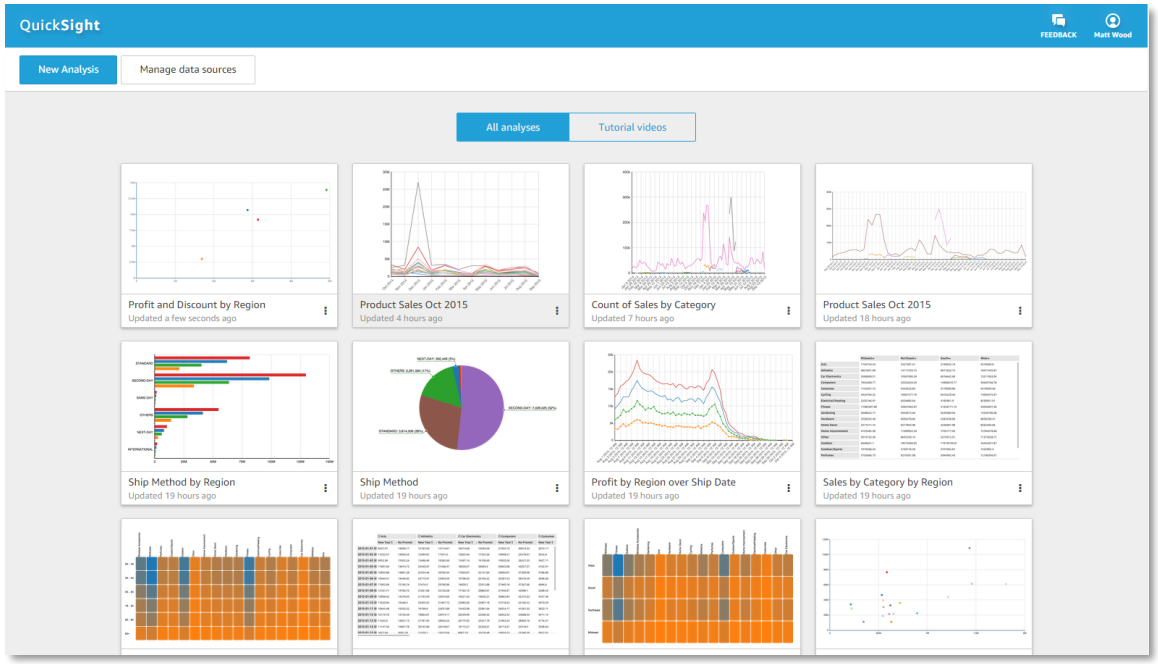
Apache Hive Metastore

Amazon QuickSight

QuickSight – fast access to all your analytics

Business user

Sign-in

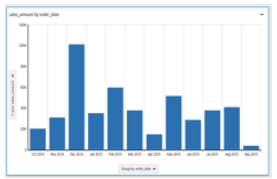


BI Analyst

Application User

Data Scientist / Engineer

Mobile Devices



Web Browsers

QuickSight UI

tableau SOFTWARE

TIBCO

Qlik

DOMO

Partner BI products

QuickSight API

Connectors

Data Prep

Metadata

SPICE

Suggestions

Files

Amazon Redshift

Amazon RDS

Amazon S3

Amazon Kinesis

Amazon DynamoDB

Amazon EMR

Apps

Direct connect JDBC/ODBC

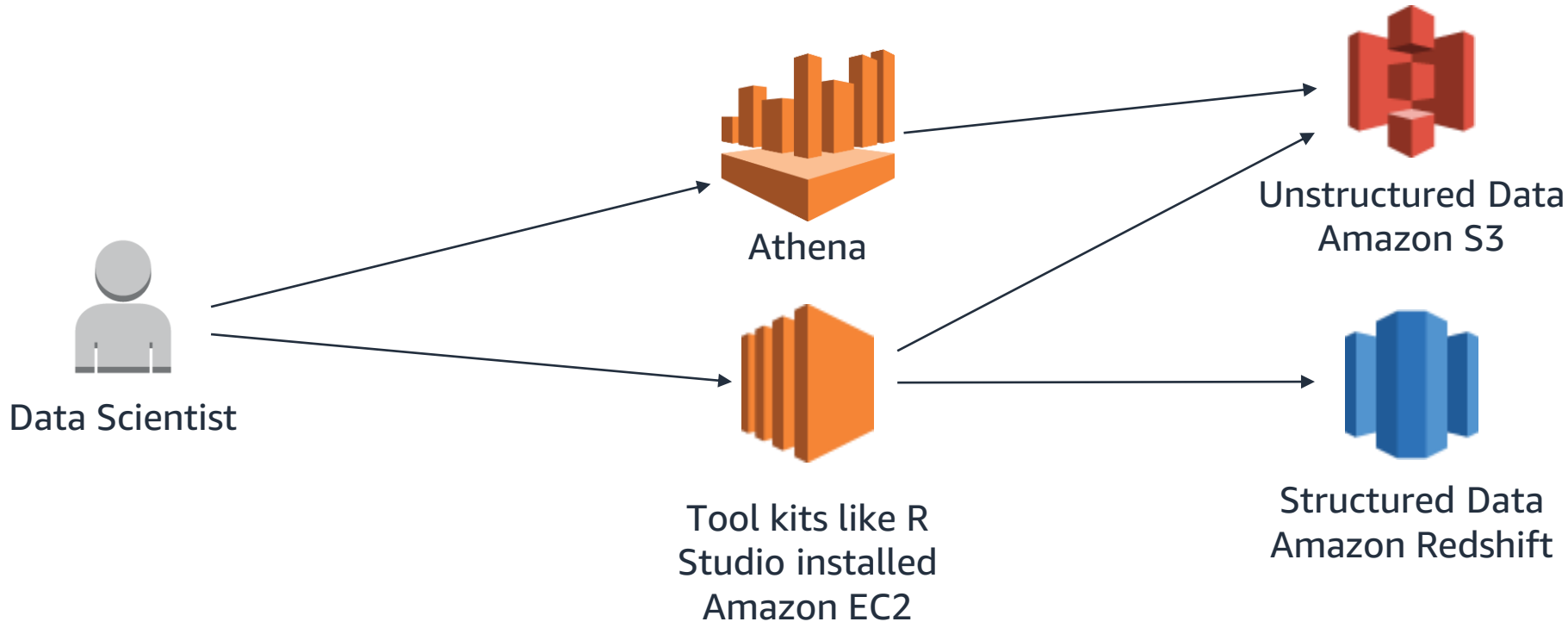
On premises Data

aws

intel

Data Scientist

Data Scientist with existing toolsets



- Work with unstructured datasets
- Use existing toolsets to connect to Redshift

What is R?

Open source programming language and software environment designed for statistical computing, data analysis, and visualization

RStudio Open source IDE for R

Shiny Server - Visualization R package for creating interactive dashboards

The New York Times

Business Computing

Search All NYTimes.com Go

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION | ARTS | STYLE | TRAVEL | JOBS | REAL ESTATE | AUTOS

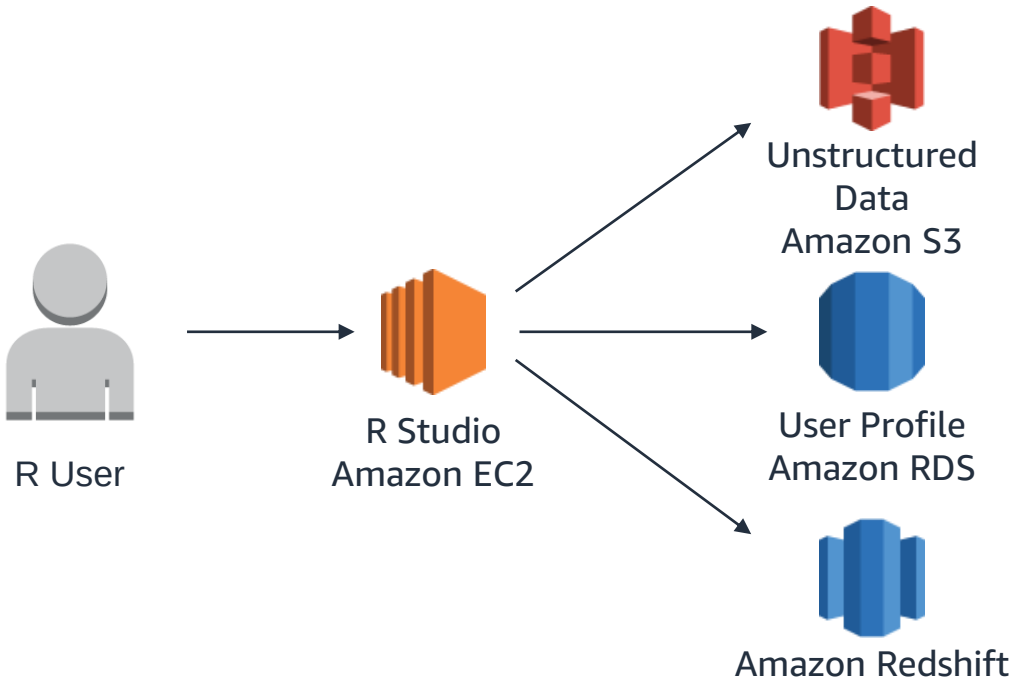
Data Analysts Captivated by R's Power

Left, Stuart Isett for The New York Times; right, Kieran Scott for The New York Times

R first appeared in 1996, when the statistics professors Robert Gentleman, left, and Ross Ihaka released the code as a free software package.

By ASHLEE VANCE
Published: January 6, 2009

Querying Redshift with R Packages



- **RJDBC** – supports SQL queries
- ***dplyr*** – Uses R code for data analysis
- **RPostgreSQL** - R compliant driver or Database Interface (DBI)

Connecting R with Amazon Redshift blog post:
https://blogs.aws.amazon.com/bigdata/post/Tx1G8828SPGX3PK/Connecting-R-with-Amazon-Redshift

Introducing Amazon Athena

Amazon Athena is an interactive query service that makes it easy to analyze data directly from Amazon S3 using Standard SQL

Query Data Directly from Amazon S3

- No loading of data
- Query data in its raw format
  - Text, CSV, JSON, weblogs, AWS service logs
  - Convert to an optimized form like ORC or Parquet for the best performance and lowest cost
- No ETL required
- Stream data from directly from Amazon S3
- Take advantage of Amazon S3 durability and availability

Use ANSI SQL

- Start writing ANSI SQL
- Support for complex joins, nested queries & window functions
- Support for complex data types (arrays, structs)
- Support for partitioning of data by any key
  - (date, time, custom keys)
  - e.g., Year, Month, Day, Hour or Customer Key, Date

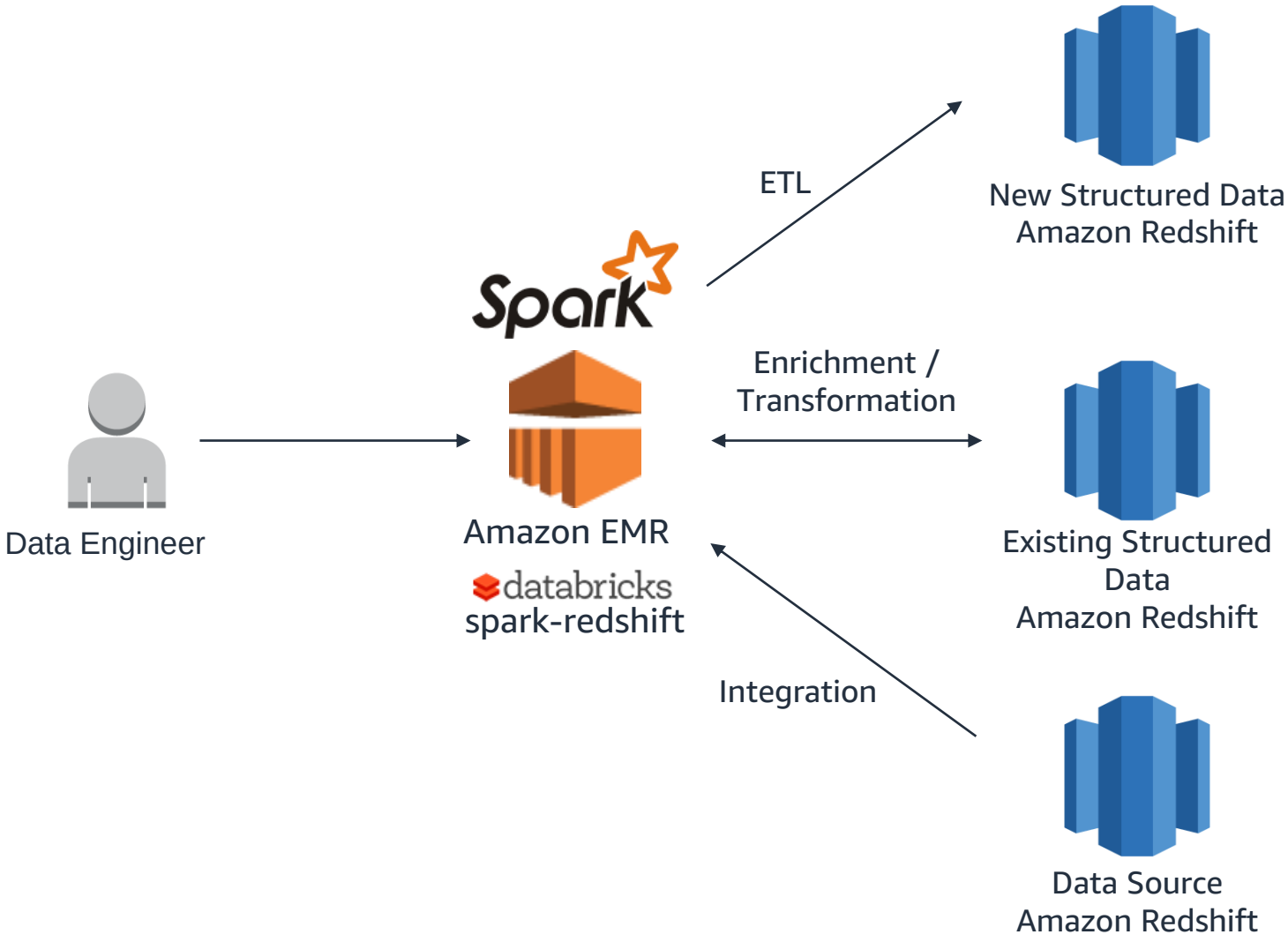
```
WITH q21_tmp1_cached AS
 (SELECT l_orderkey,
         count(DISTINCT l_suppkey) AS count_suppkey,
         max(l_suppkey) AS max_suppkey
  FROM lineitem_parq
  WHERE l_orderkey IS NOT NULL
  GROUP BY l_orderkey),
    q21_tmp2_cached AS
 (SELECT l_orderkey,
         count(DISTINCT l_suppkey) count_suppkey,
                                   max(l_suppkey) AS max_suppkey
  FROM lineitem_parq
  WHERE l_receiptdate > l_commitdate
    AND l_orderkey IS NOT NULL
  GROUP BY l_orderkey)
SELECT s_name,
       count(1) AS numwait
FROM
  (SELECT s_name
   FROM
     (SELECT s_name,
             t2.l_orderkey,
             l_suppkey,
             count_suppkey,
             max_suppkey
      FROM q21_tmp2_cached t2
      RIGHT OUTER JOIN
        (SELECT s_name,
                l_orderkey,
                l_suppkey
         FROM
           (SELECT s_name,
                   t1.l_orderkey,
                   l_suppkey,
                   count_suppkey,
                   max_suppkey
            FROM q21_tmp1_cached t1
            JOIN
              (SELECT s_name,
                      l_orderkey,
                      l_suppkey
               FROM orders_parq o
               JOIN
                 (SELECT s_name,
                         l_orderkey,
                         l_suppkey
                  FROM nation_parq n
                  JOIN supplier s ON s.s_nationkey = n.n_nationkey
                  AND n.n_name = 'SAUDI ARABIA'
                  JOIN lineitem_parq l ON s.s_suppkey = l.l_suppkey
                  WHERE l.l_receiptdate > l.l_commitdate
                    AND l.l_orderkey IS NOT NULL) l1 ON o.o_orderkey = l1.l_orderkey
               AND o.o_orderstatus = 'F') l2 ON l2.l_orderkey = t1.l_orderkey) a
           WHERE (count_suppkey > 1)
             OR ((count_suppkey=1)
                 AND (l_suppkey <> max_suppkey))) l3 ON l3.l_orderkey = t2.l_orderkey) b
   WHERE (count_suppkey IS NULL)
     OR ((count_suppkey=1)
         AND (l_suppkey = max_suppkey))) c
GROUP BY s_name
ORDER BY numwait DESC,
         s_name LIMIT 100;
```

Amazon Athena is Cost Effective

- Pay per query
- $5 per TB scanned from S3
- DDL Queries and failed queries are free
- Save by using compression, columnar formats, partitions

Data Engineer

Data Engineer familiar with Hadoop and Spark



Spark SQL + DataFrames

Streaming

MLlib
Machine Learning

GraphX
Graph Computation

Spark Core API

R

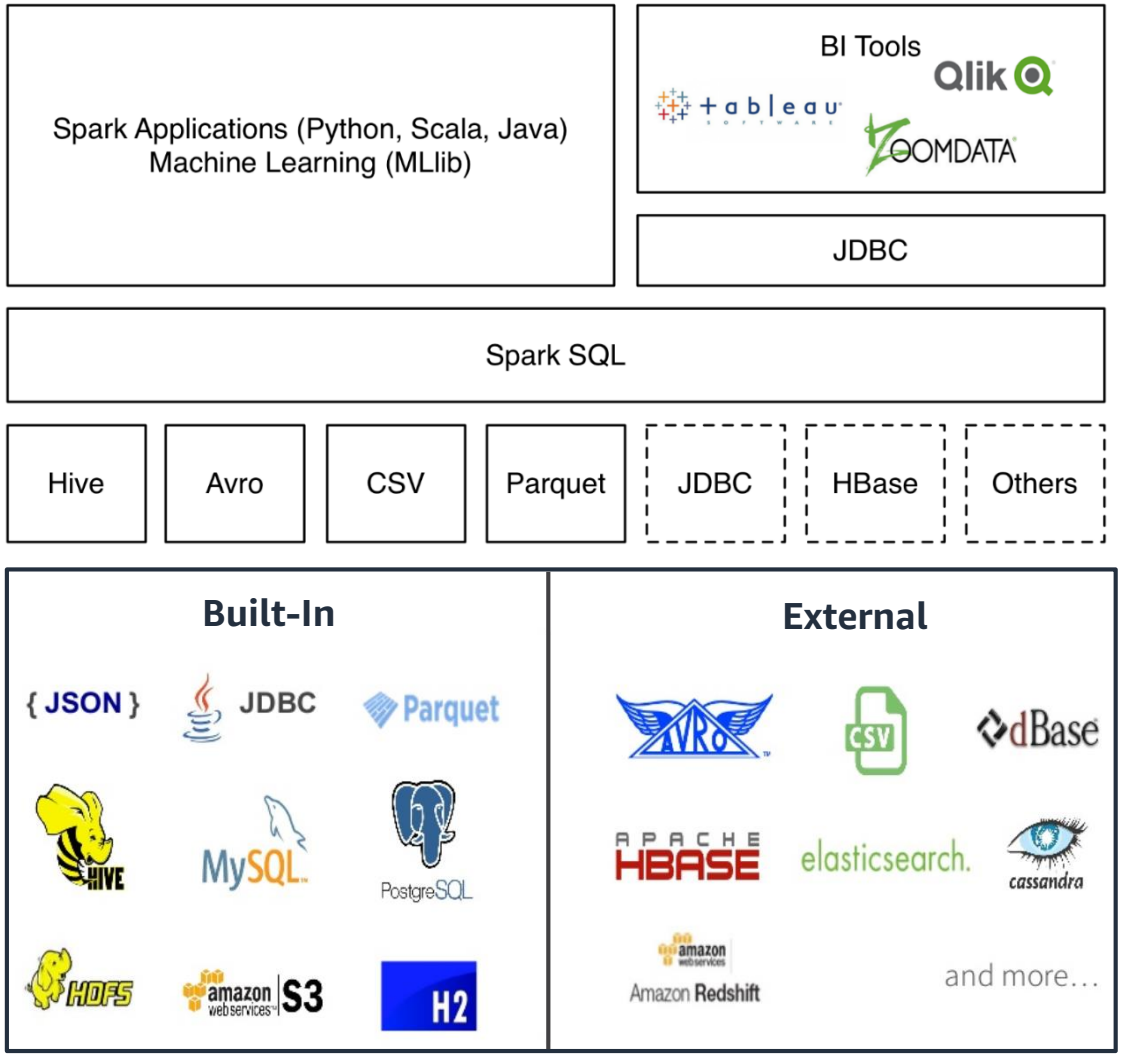
SQL

Python

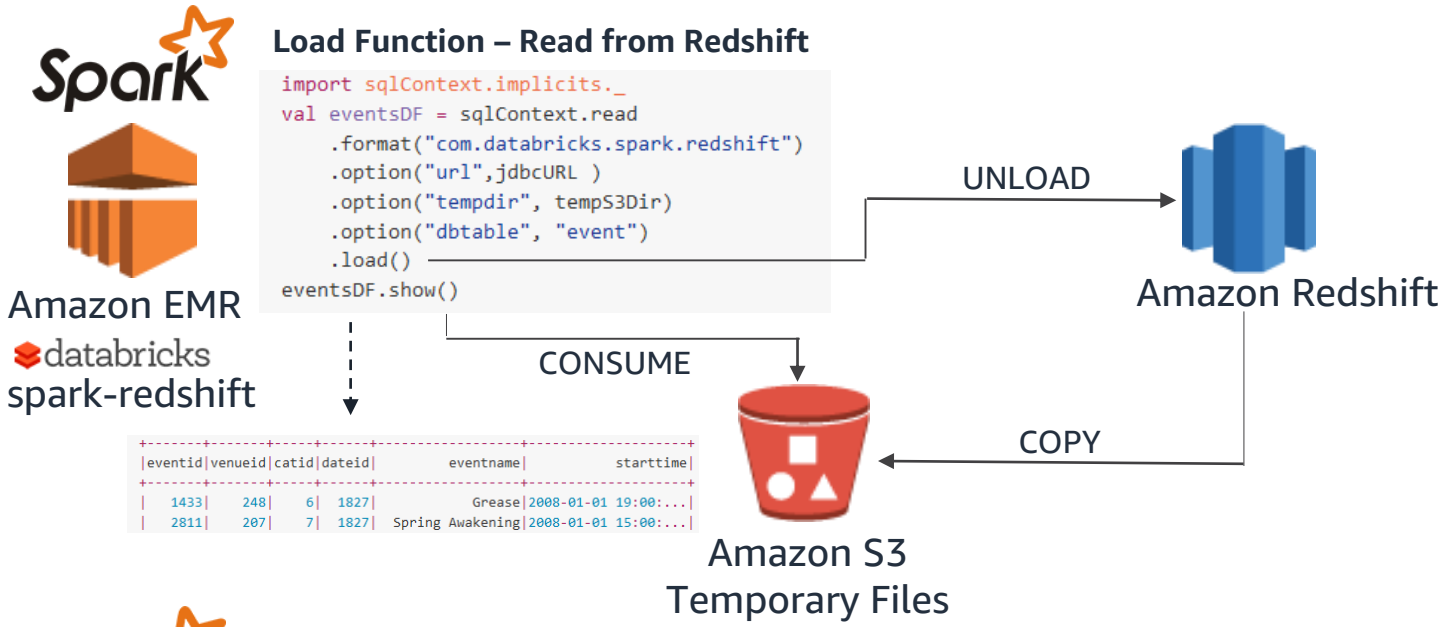
Scala

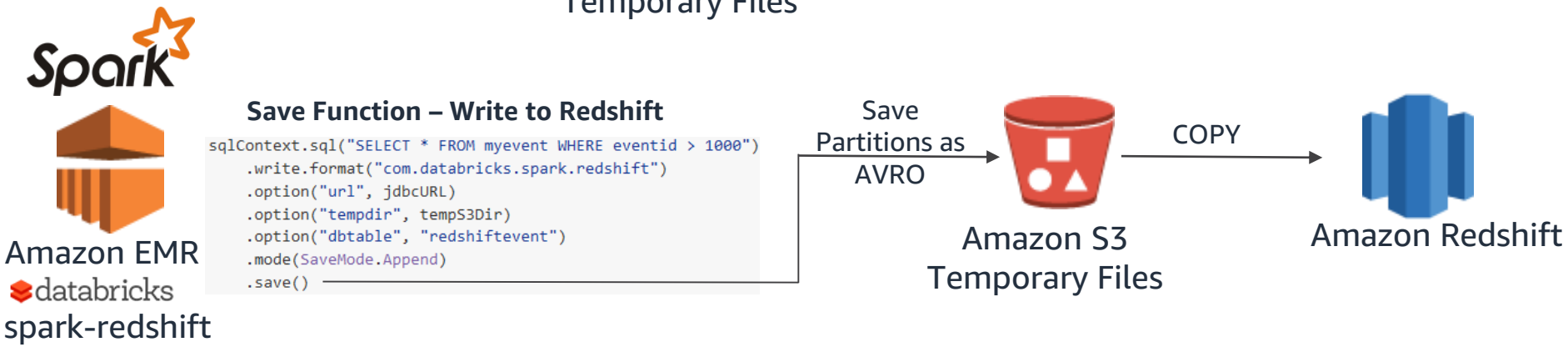
Java

Spark SQL Data Sources API



Manipulating Redshift with spark-redshift





Databricks Tutorial: https://github.com/databricks/spark-redshift/tree/master/tutorial

Learn from AWS experts. Advance your skills and knowledge. Build your future in the AWS Cloud.

Digital Training

Free, self-paced online courses built by AWS experts

Classroom Training

Classes taught by accredited AWS instructors

AWS Certification

Exams to validate expertise with an industry-recognized credential

Ready to begin building your cloud skills?
Get started at: https://www.aws.training/

With deep expertise on AWS, APN Partners can help your organization at any stage of your Cloud Adoption Journey.

AWS Managed Service Providers

APN Consulting Partners who are skilled at cloud infrastructure and application migration, and offer proactive management of their customer's environment.

AWS Competency Partners

APN Partners who have demonstrated technical proficiency and proven customer success in specialized solution areas.

AWS Marketplace

A digital catalog with thousands of software listings from independent software vendors that make it easy to find, test, buy, and deploy software that runs on AWS.

AWS Service Delivery Partners

APN Partners with a track record of delivering specific AWS services to customers.

Ready to get started with an APN Partner?
Find a partner: https://aws.amazon.com/partners/find/
Learn more at the AWS Partner Network Booth

Thank You for Attending AWS Innovate

We hope you found it interesting! A kind reminder to **complete the survey.**

Let us know what you thought of today's event and how we can improve the event experience for you in the future.


aws-apac-marketing@amazon.com


twitter.com/AWSCloud


facebook.com/AmazonWebServices


youtube.com/user/AmazonWebServices


slideshare.net/AmazonWebServices


twitch.tv/aws