AWS Neptune

Abhijit Gokhale
Shubham Sharma

TABLE OF CONTENTS

01 **CHALLENGES**
Problems with Relational and other Graph databases

02 **WHY AWS Neptune?**
What does AWS Neptune provide?

03 **USE CASES**
Where Graph database can be used?

04 **GRAPH MODELS**
How to store, retrieve, and analyze data

05 **COMPETITORS**
Azure, Google Cloud

06 **DEMO**
What data have we created?

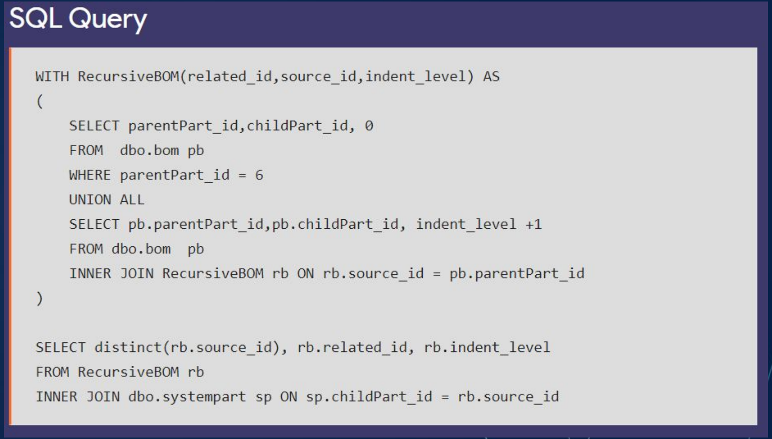
01

CHALLENGES

Problems with Relational and why use Graph Databases

Relational Database

- Unnatural for Graph Querying
- Inefficient graph processing
- Rigid schema inflexible for changing data
- Have to write large number of joins for graph traversal which can become too slow



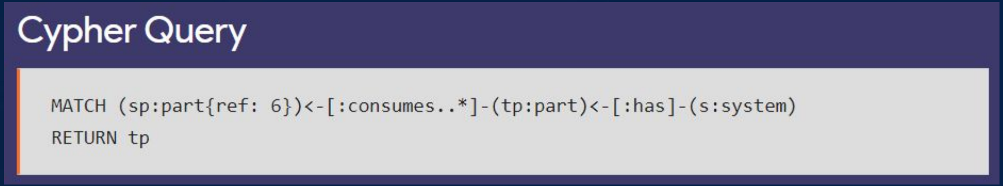
SQL Query

```
WITH RecursiveBOM(related_id,source_id,indent_level) AS
(
    SELECT parentPart_id,childPart_id, 0
    FROM  dbo.bom pb
    WHERE parentPart_id = 6
    UNION ALL
    SELECT pb.parentPart_id,pb.childPart_id, indent_level +1
    FROM dbo.bom  pb
    INNER JOIN RecursiveBOM rb ON rb.source_id = pb.parentPart_id
)

SELECT distinct(rb.source_id), rb.related_id, rb.indent_level
FROM RecursiveBOM rb
INNER JOIN dbo.systempart sp ON sp.childPart_id = rb.source_id
```

Why use Graph Database

- Connections or relationships between entities are at the core of the data
- Easy to model data interconnections as a graph, and then write complex queries that extract real-world information
- Advantageous over complex relational databases
- A graph database like Neptune can query relationships between billions of vertices without bogging down.



Cypher Query

```
MATCH (sp:part{ref: 6})<-[:consumes..*]-(tp:part)<-[:has]-(s:system)
RETURN tp
```

02

WHY AWS Neptune ?

KEY FEATURES

- High performance and Scalability
- High Availability and Durability
- Query Support - Apache TinkerPop Gremlin, the W3C’s SPARQL, and Neo4j's openCypher
- Highly Secure
- Fully Managed
- Fast Parallel Bulk Data Loading
- Cost Effectiveness - Pay only for what you use

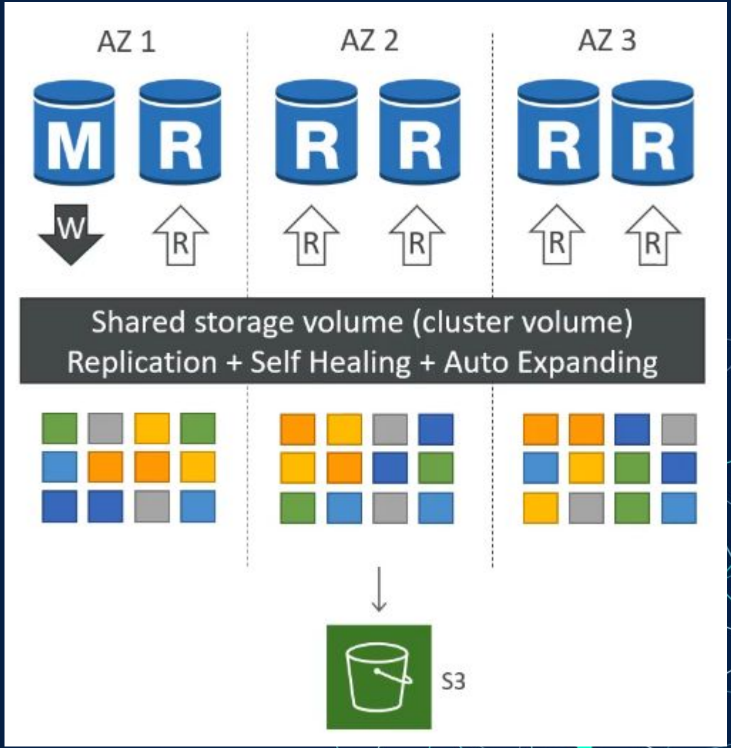
ARCHITECTURE

- Fault Tolerance achieved by replicating data 6 times across 3 availability zones. One instance is the master. ACID model with immediate consistency

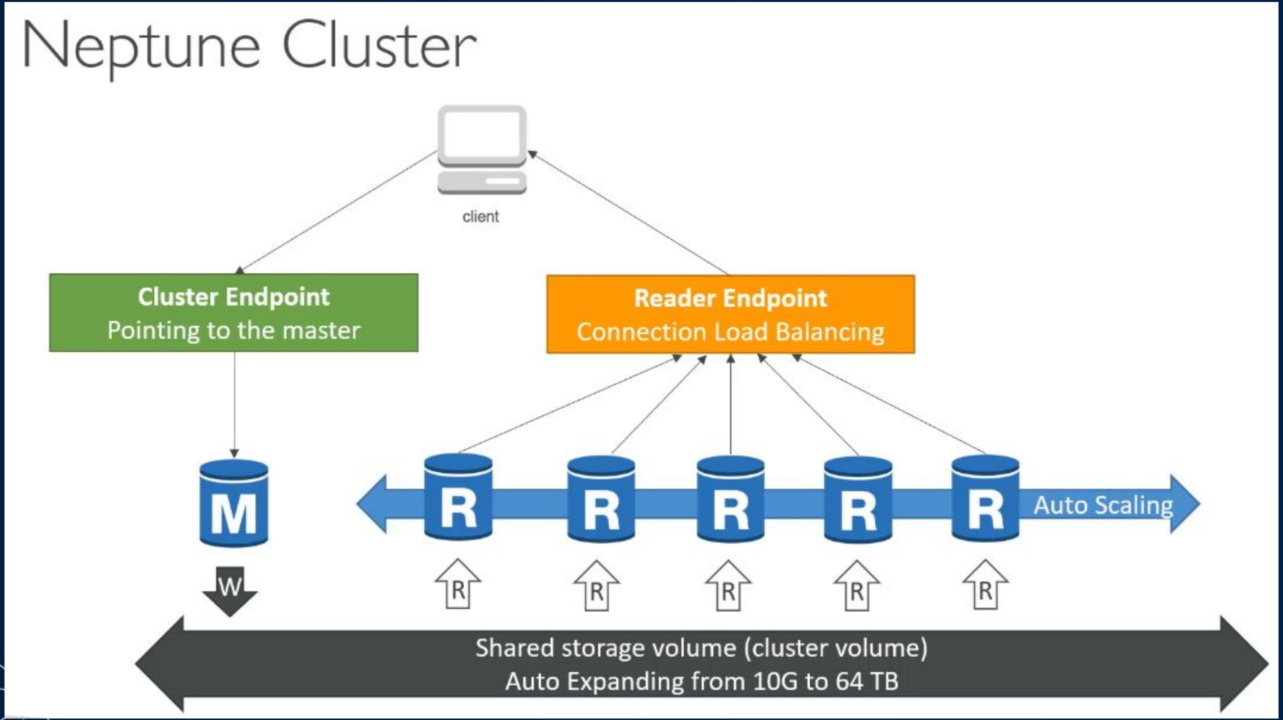
- Storage is striped across 100s of volumes with each being 10GB

- Data is stored using Lock-free optimistic algorithm. Data is considered durable when at least 4/6 copies acknowledge the write. For read, it uses 3/6 quorum model

- Data is continuously backed up to S3 in real time, using storage nodes


AZ 1
AZ 2
AZ 3
M
R
R
R
R
R
W
R
R
R
R
R
Shared storage volume (cluster volume)
Replication + Self Healing + Auto Expanding
S3

What a cluster looks like



03

USE CASES

Social Networking

Easily process social network information over large sets of user profiles and interactions. Used to find common interaction in Twitter, Facebook

Existing recommender system methods use metrics of similarity to recommend other nodes which do not take into account the graph structure of the relationships between the nodes.

Recommendations

Fraud Detection

Using a fraud graph, organizations can identify a network of connected users and items such as e-mail accounts, addresses, and phone numbers that they have in common.

Use graph-based technique for data integration, management of research publications, drug discovery, precision medicine, and cancer research

Life Sciences

04

GRAPH DATA MODEL

PROPERTY GRAPH

TinkerPop

Open Source package maintained by Apache

Gremlin

Provides a universal Graph Query Language, Gremlin which is easy to integrate with Java and Python

Nodes and Edges

Nodes and Edges are built connecting each other where both of them can have some attributes

RESOURCE DESCRIPTION FRAMEWORK

W3C

Standard Formats specified by World Wide Web Consortium (W3C)

5 Formats specified

Intended for situations in which information on the Web needs to be processed by applications

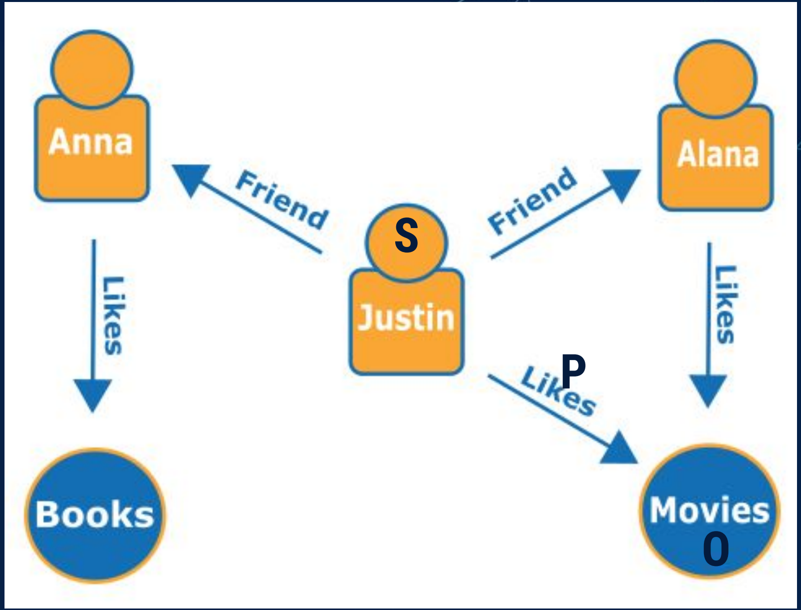
SPARQL

Mostly like SQL consisting of SELECT, UPDATE, INSERT, WHERE clauses

RESOURCE DESCRIPTION FRAMEWORK

- The subjects and objects of the triples make up the nodes in the graph; the predicates form the arcs.

- Below is an example of “**triples**”
  <Bob> <is a> <person>.
  <Bob> <is a friend of> <Alice>.

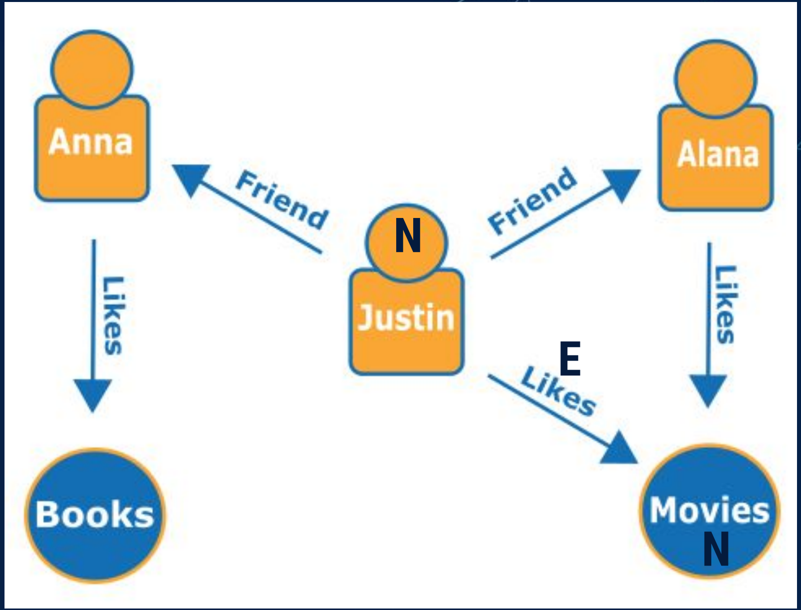


GREMLIN

- Nodes and edges define the graph which is similar to openCypher

- Visualization is difficult in Gremlin. It is mostly for traversing the graph

- Compatible with both OLTP and OLAP engines



05

COMPETITORS

Microsoft Azure Cosmos DB

- Azure Cosmos DB offers document, graph, key-value and wide-column data models in a single service. If our requirement is a multi-model database, **Azure will be cheaper**

- Azure Cosmos DB does not support RDF SPARQL data models. It only supports Gremlin.

Google Cloud Platform | Neo4j AuraDB

- In 2020, Neo4j announced availability of Neo4j AuraDB on Google Cloud as the only integrated graph database service on GCP

- Amazon Neptune does not support advanced data analytics with solutions such as Spark and GraphX

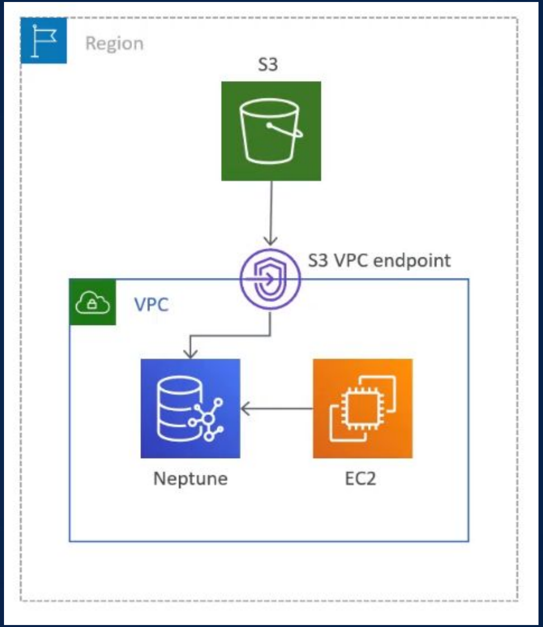
- In Neo4j, updates are typically made from the master which has no regard for the number of instances that fail as long as it remains available whereas AWS Neptune uses quorum model where certain replicas have to acknowledge reads/writes

06

DEMO

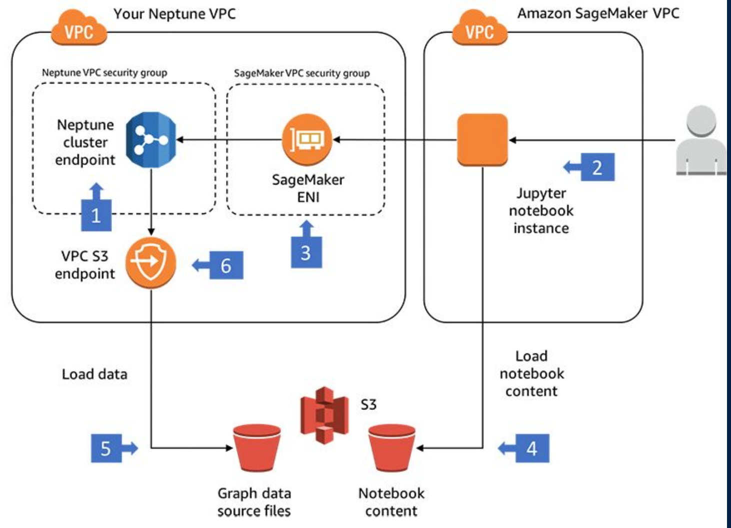
Bulk Load Data from S3 - Using EC2

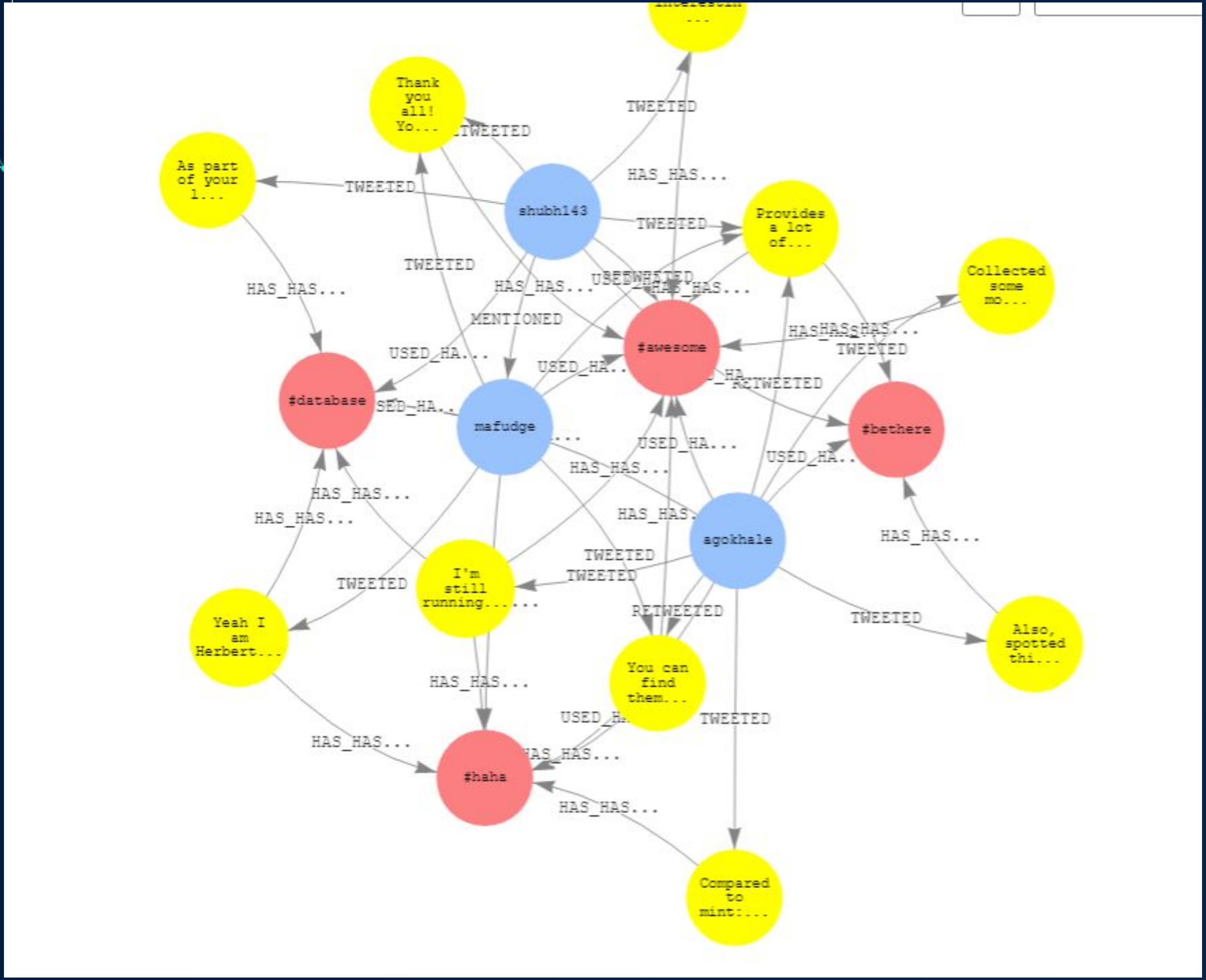
- We perform an HTTP Post request to the loader endpoint
- Neptune Cluster must assume an IAM role with S3 read access
- For Gremlin, we can upload CSV files
- For SPARQL, we can upload different file formats - ntripples, nquads, rdfxml, turtle



Bulk Load Data from S3 - Jupyter Notebook

- We use magic functions that begin with “%” or “%%”
- Need to specify S3 Bucket, Amazon Resource Name (ARN) and Region
- ARN allows Neptune Cluster to connect to S3




Interestin
...
Thank you all! Yo...
As part of your l...
shubh143
Provides a lot of...
Collected some mo...
#awesome
#database
mafudge
#bethere
agokhale
I'm still running...
Yeah I am Herbert...
You can find them...
Also, spotted thi...
#haha
Compared to mint:...
TWEETED
RETWEETED
HAS_HAS...
USED_HA...
MENTIONED

References

- https://aws.amazon.com/neptune/getting-started/
- https://aws.amazon.com/neptune/features/
- https://docs.aws.amazon.com/neptune/latest/userguide/intro.html
- https://aws.amazon.com/blogs/database/building-a-customer-identity-graph-with-amazon-neptune/
- https://www.zdnet.com/article/aws-neptune-update-machine-learning-data-science-and-the-future-of-graph-databases/
- https://db-engines.com/en/system/Amazon+Neptune%3BMicrosoft+Azure+Cosmos+DB
- https://www.peerspot.com/products/comparisons/amazon-neptune_vs_microsoft-azure-cosmos-db
- https://stackshare.io/stackups/amazon-neptune-vs-neo4j
- https://leapgraph.com/aws-neptune-vs-neo4j/

THANK YOU