SyncOasis
Bring Your Information into Sync

Analytic Data Stores: Should I Really Use a Database?

November 13, 2018

Intro

Who are we?

What is this session about?

1. **You will know when to use databases over Hadoop, which is vast majority of use cases**
2. **You will be able to choose the right database for your use case, since databases are no longer a "commodity"**
3. **You will be able to model traditional data into a Document Databases**

Do you represent any vendors?

Intro Poll

Q1. Hadoop is:
- a) A filesystem that shards data across commodity hardware
- b) A database that shards data across commodity hardware
- c) A new database for the cloud
- d) A new twist on the hula hoop

Q2. Parquet is:
- a) A new floor style
- b) A data storage format in Hadoop
- c) A query language for Hadoop
- d) A new database offering in AWS

Intro Poll

Q3. JSON is:
- a) Just Some Other Noobie
- b) JavaScript Object Notation
- c) Java Standard Object Notation
- d) Jupyter Served On Network

Q4. Which type of data format includes markup?
- a) XML
- b) JSON
- c) Tables
- d) Avro

Intro Poll

Q5. What are NoSQL databases?
- a) Databases that aren't SQL Server
- b) Non-mainstream databases
- c) New relational databases that query data with another language
- d) Databases that do not store data in traditional tables

Q6. What is true about relational databases?
- a) Only organize data by row within a table
- b) Generally scale horizontally
- c) Generally scale vertically
- d) Do not offer features found in other types of databases

Intro Poll

Q7. Which statement is true?
- a) A column-oriented database and a wide column store are the same
- b) All databases of a certain type generally have the same features
- c) There are standard query languages for NoSQL databases like SQL
- d) JavaScript is a typical language used with NoSQL databases

Q8. Document Databases store data primarily in:
- a) Excel documents
- b) Word documents
- c) JSON documents
- d) ORC documents

What's the difference?

- Data Storage
- Compute

- Data Storage
- Compute

Hadoop logo courtesy of Apache Software Foundation and displayed under Apache License, Version 2.0

Hadoop Particulars

Data Storage

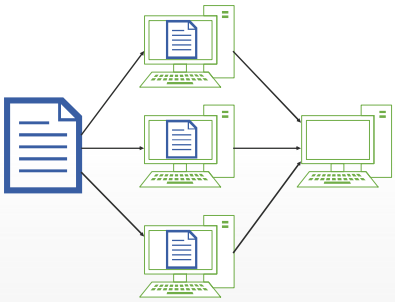
- Hadoop Distributed Files System (HDFS)

Compute

- MapReduce
- Spark

Key takeaway: Both data storage and compute are distributed

MapReduce



MapReduce Shortcomings

It can be really slow!
It must be possible to split the computation into parts!

MapReduce Alternative?

And yet

- Requires gobs of memory
- Easy to overwhelm servers
- Isn't great with small files

Much faster than MapReduce!

Spark logo courtesy of Apache Software Foundation and displayed under Apache License, Version 2.0

Hadoop Poll

Q1. Which of the following is not associated with Hadoop:
a) MapReduce
b) HDFS
c) WiredTiger
d) Spark

Q2. Hadoop is:
a) A database
b) A distributed file and compute system
c) Runs on specialized hardware
d) Stores data in many small files

Q3. Hadoop works best for:
a) Massively large (100s TB+) sets of data
b) Quick aggregations of data
c) Quick manipulation of data
d) Creating organized data for analytics

Case Study - Why a Database over Hadoop?

Data

- Several main data sets with additional lookup tables
- 100,000s of new entries daily

Uses

- Perform CRUD on the data
- Run reports, perform analysis

Key takeaway: Your data should be more than just a parking lot of static data, it should be a living / breathing data asset that is managed and curated.

Case Study

- Clean Data
- Train ML Model

- Copy from Hadoop
- Clean Data
- Train ML Model
- Test ML Model
- Predict on new data
- Copy back to Hadoop

Hadoop vs Databases

- UD not generally supported
- Joining non-trivial
- Supports ***HUGE*** data
- Supports columnar data

- CRUD fully supported
- Joining fully supported
- Can support ***HUGE*** data
- Can support columnar data

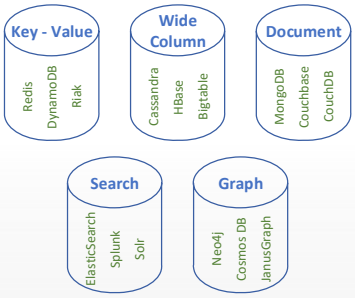
Do you really have Petabytes of data?

Case Study Discussion

Key takeaways

- Databases are often the right approach, the key is choosing the right database for your needs
- Column based versus row based tables are usually superior for analytic use cases

NoSQL Database Types



Key-Value Database

Table Version

RestaurantID	Name	Borough	Cuisine
30075445	Morris Park Bake Shop	Bronx	Bakery
30112340	Wendy's	Brooklyn	Hamburgers

Key-Value Version

"Restaurant-30075445-Name": "Morris Park Bake Shop"
"Restaurant-30075445-Borough": "Bronx"
"Restaurant-30075445-Cuisine": "Bakery"
"Restaurant-30112340-Name": "Wendy's"
"Restaurant-30112340-Borough": "Brooklyn"
"Restaurant-30112340-Cuisine": "Hamburgers"

Wide Column Store

Restaurant Table:

RestaurantID	Name	Borough	Cuisine
30075445	Morris Park Bake Shop	Bronx	Bakery
30112340	Wendy's	Brooklyn	Hamburgers

Address Table:

RestaurantID	AddressID	ZipCode	Longitude	Latitude
30075445	1	10462		
30112340	2		-73.961704	40.662942

Wide Column Store Table:

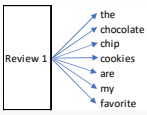
	Borough	Cuisine	ZipCode
Morris Park Bake Shop	Bronx	Bakery	10462

	Borough	Cuisine	Longitude	Latitude
Wendy's	Brooklyn	Hamburgers	-73.961704	40.662942

Search Engine

Reviews Table

RestaurantID	ReviewID	Date	Review
30075445	1	1/24/2013	The chocolate chip cookies are my favorite
30075445	2	9/11/2013	I love the chocolate truffle cake
30112340	3	4/17/2014	My favorite burgers



	Review1	Review2	Review3
are	x		
burgers			x
cake		x	
chip	x		
chocolate	x	x	
cookies	x		
favorite	x		x
i		x	
love		x	
my	x		x
the	x	x	
truffle		x	

Graph Database

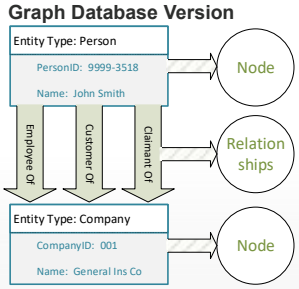
Table Version

EmployeeID	Person	CompanyID
1001	John Smith	001

CustomerID	Person	CompanyID
1000-1572	John Smith	001

ClaimantID	Person	CompanyID
PN0001-003	John Smith	001

Graph Database Version



Document Database

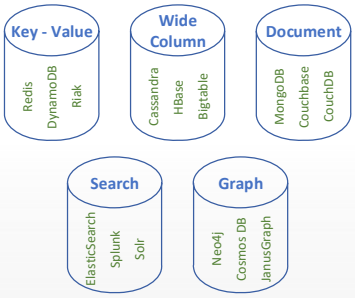
Table Version

EmployeeID	Person	HireDate
1001	John Smith	1/1/2015

JSON Version

```
{
  "EmployeeID": 1001,
  "Person": "John Smith",
  "HireDate": 2015-01-01
}
```


NoSQL Database Recap



Notes – NoSQL Database Recap

Key takeaways

- Need to choose the right NoSQL database type
- Within each database type, product capabilities can vary significantly by vendor
- Choosing right database is important, as there is no universal "SQL"
- A mixed model approach may be appropriate (i.e. a database that uses more than one NoSQL database type)
- Storing the data "twice" may be appropriate (within a database, between database types)
- How you store the data can be different from how you use the data

Notes – NoSQL Database Use Cases

Key-Value Database
- Store session data for live applications, store lookups
- Real-time ingest layer for other databases

Wide Column Store
- Massive ingestion of data (i.e. IOT / telematics)
- A two-dimensional key value store (i.e. organize key-values into a record)

Search Engine
- Find records that match criteria without writing SQL
- Find information in unstructured fields, since every word is indexed

Graph Database
- Fraud Detection, Recommendation Engine, Master Data Management, Social Network Graphs, AI and ML

Document Database
- Customer 360 / Siloed data integration
- Flexible schema within a structured data model
- Store hierarchical data or data with variable number of fields

Employee Data

Employee Number:	1001-001-0589
Employee Name:	Jane Smith
Job Title:	Data Engineer
Business Address:	1001 Cherry Way, Springfield, MO 65807
Home Address:	989 Arroyo Street, Springfield, MO 65801
Date of Birth:	XX/XX/XXXX
SSN:	XXX-XX-XXXX
Programming Languages:	SQL, SAS, JavaScript, Python, Scala, Go

Employee Data in Tables

EmployeeID	Name	Title	Date of Birth	SSN
1001-001-0589	Jane Smith	Data Engineer	XX/XX/XXXX	XXX-XX-XXXX

EmployeeID	AddressID	Type	Street	City	State	ZipCode
1001-001-0589	001	Business	1001 Cherry Way	Springfield	MO	65807
1001-001-0589	002	Home	989 Arroyo Street	Springfield	MO	65801

EmployeeID	LanguageID
1001-001-0589	001
1001-001-0589	002
1001-001-0589	003
1001-001-0589	004
1001-001-0589	005
1001-001-0589	006

LanguageID	Name	VM
001	SQL	Database
002	SAS	SAS
003	JavaScript	Browser
004	Python	PVM
005	Scala	JVM
006	Go	N/A

Document Objects

Array

```
{
  "Programming Languages": ["SQL", "SAS", "JavaScript", "Python", "Scala", "Go"]
}
```

Sub-Object

```
{
  "Address": {
    "Type": "Business",
    "Street": "1001 Cherry Way",
    "City": "Springfield",
    "State": "MO",
    "ZipCode": "65807"
  }
```


Employee Data in Document

```
{
  "EmployeeNumber": "1001-001-0589",
  "EmployeeName": "Jane Smith",
  "JobTitle": "Data Engineer",
  "Address": [
    { "Type": "Business", "Street": "1001 Cherry Way",
      "City": "Springfield", "State": "MO", "ZipCode": "65807" },
    { "Type": "Home", "Street": "989 Arroyo Street",
      "City": "Springfield", "State": "MO", "ZipCode": "65801" }
  ],
  "Date of Birth": XXXX-XX-XX,
  "SSN": "XXX-XX-XXXX",
  "Programming Languages": ["SQL", "SAS", "JavaScript", "Python", "Scala", "Go"]
}
```


Document Exercise

Property Analysis

AY	12	24	36	48	60
2013	100	125	135	138	138
2014	150	185	220	225	
2015	135	190	200		
2016	175	235			
2017	140				

Factors	12-24	24-36	36-48	48-Ult
2013	1.250	1.080	1.022	1.000
2014	1.233	1.189	1.023	
2015	1.407	1.053		
2016	1.343			
Avg	1.308	1.107	1.022	1.000
Wtd	1.313	1.110	1.023	1.000
Selected	1.310	1.110	1.025	1.000
CDF	1.490	1.138	1.025	1.000

Methods

AY	Cur	CDF	LDF Ult	EP	LR
2013	138	1.000	138	200	69.0%
2014	225	1.000	225	300	75.0%
2015	200	1.025	205	350	58.6%
2016	235	1.138	267	400	66.8%
2017	140	1.490	209	375	55.7%

AY	ELR	Unpaid	BF Ult	Sel Ult
2013	60%	0%	138	138
2014	60%	0%	225	225
2015	60%	2%	205	205
2016	60%	12%	264	265
2017	60%	33%	214	215

Document Exercise "Answer"

See the relevant article on my LinkedIn feed:

www.linkedin.com/in/jeff-white-syncoasis/detail/recent-activity/posts