**DE Scenarios – Week 8**

What technologies would you use to gather, process, store, search, and report for these scenarios?

**Options**:

**Categories**: Key Value, Document, Graph, Columnar, traditional, dw

**Data Stores**: Redis, Hadoop, Snowflake, RDBMS, Hbase, Cassandra, MongoDB

**Process Engines**: Pig, Hive, Spark, Storm, MapReduce, Python, ML, query editor

**Reporting**: Python, Power BI, Splunk, ML, etc.

Questions to ask:

What is the nature/kind of data

How would you store that kind of data

What do you intend to do with it

1. Archive millions of emails. Remove sensitive data (passwords, social security numbers, political bribes…)

* Category:
  + Document
  + Columnar
* Storage:
  + MongoDB
  + Cassandra (column families)
* Processing:
  + MongoDB
* Reporting:
  + Power BI
  + Splunk

1. Read through decades of hospital employee data to see if vocabulary changes predict onset of dementia.

* Category:
  + Document
  + Key-value (video)
* Storage:
  + MongoDB
  + Hadoop
* Processing:
  + MongoDB
  + Spark
  + Snowflake
* Reporting:
  + ML

1. Report team needs to mine years of log files for traffic spike patterns

* Category:
  + Document
* Storage:
  + MongoDB
* Processing:
  + MongoDB
* Reporting:
  + Splunk

1. Monitor tweets and alert me when it is finally time to bring back disco.

* Category:
  + Key-value
* Storage:
  + Storm
* Processing:
  + Storm
* Reporting:
  + Power BI

1. Search through terabytes of videos for any that mention your company or display your logo.

* Category:
  + Key-Value
  + Document
* Storage:
  + Hadoop
  + MongoDB
* Processing:
  + Spark
  + Pig
  + MongoDB
* Reporting:
  + ML

1. Thousands of contracts in Word need to be quickly searchable by date, contract type, amount, and customer

* Category:
  + Columnar
  + Document
* Storage:
  + Hbase/Cassandra
  + MongoDB
* Processing:
  + Hive
* Reporting:
  + splunk

1. Predict changes in stock value from historic data

* Category:
  + Key-value
* Storage:
  + hadoop
* Processing:
  + spark
* Reporting:
  + Python/ml

1. Archive a large DB2 database before the server is decommissioned. It must be searchable if needed.

* Category:
  + traditional
* Storage:
  + Rdbms/snowflake
* Processing:
* Reporting:
  + Power BI

1. Store audio files for customers but flag any that contain copyrighted material.

* Category:
  + Key-Value
  + Document
* Storage:
  + Hadoop
  + MongoDB
* Processing:
  + Spark
  + Pig
  + MongoDB
* Reporting:
  + ML

1. Reports are hurting DB server performance. Fix it. Reporting data can be up to 24 hours old

* Category:
  + dw
* Storage:
  + snowflake
* Processing:
  + Pre-processed cube
* Reporting:
  + pbi

1. Store web transaction data. Changes are happening all the time and reports HAVE to be 100% accurate.

* Category:
  + Document
* Storage:
  + rdbms
* Processing:
  + Kafka/Storm
* Reporting:
  + pbi

1. Buffer web form data to prevent traffic spikes from overwhelming DB server

* Category:
  + traditional
* Storage:
  + rdbms
* Processing:
  + Kafka/storm
* Reporting:
  + pbi

1. Alert us of irregularities in network traffic that could indicate hacking

* Category:
  + Key-value
* Storage:
  + hadoop
* Processing:
  + Kafka/storm
* Reporting:
  + splunk

1. Read sensor data and upload it to oracle database every hour

* Category:
  + traditional
* Storage:
  + oracle
* Processing:
  + Kafka/storm
* Reporting:
  + Pbi

1. As new resume files are saved (Word, PDF, etc.) make them searchable by location, education, skills, etc.

* Category:
  + Document
* Storage:
  + MongoDB
* Processing:
  + MongoDB
* Reporting:
  + Python, splunk

1. Look through millions of photos. Find ones that have Football Logos

* Category:
  + Key-Value
  + Document
* Storage:
  + Hadoop
  + MongoDB
* Processing:
  + Spark
  + Pig
  + MongoDB
  + ml
* Reporting:
  + ML