Big Data Basics

Justin Post

Where are we?

- Foundation in Python through `JupyterLab`

Now:

- What is Big Data?
- What are common things we'll run into with Big Data?
- What issues are we going to try and tackle?
- What software can be useful?

What is Big Data?

Useful definition:

- Big data = data that you can't handle 'normally'

  - data is too large to read into memory and use our usual software
  - Streaming data, we don't have all observations at the start, how to we update our stats + models

What is Big Data?

Useful definition:

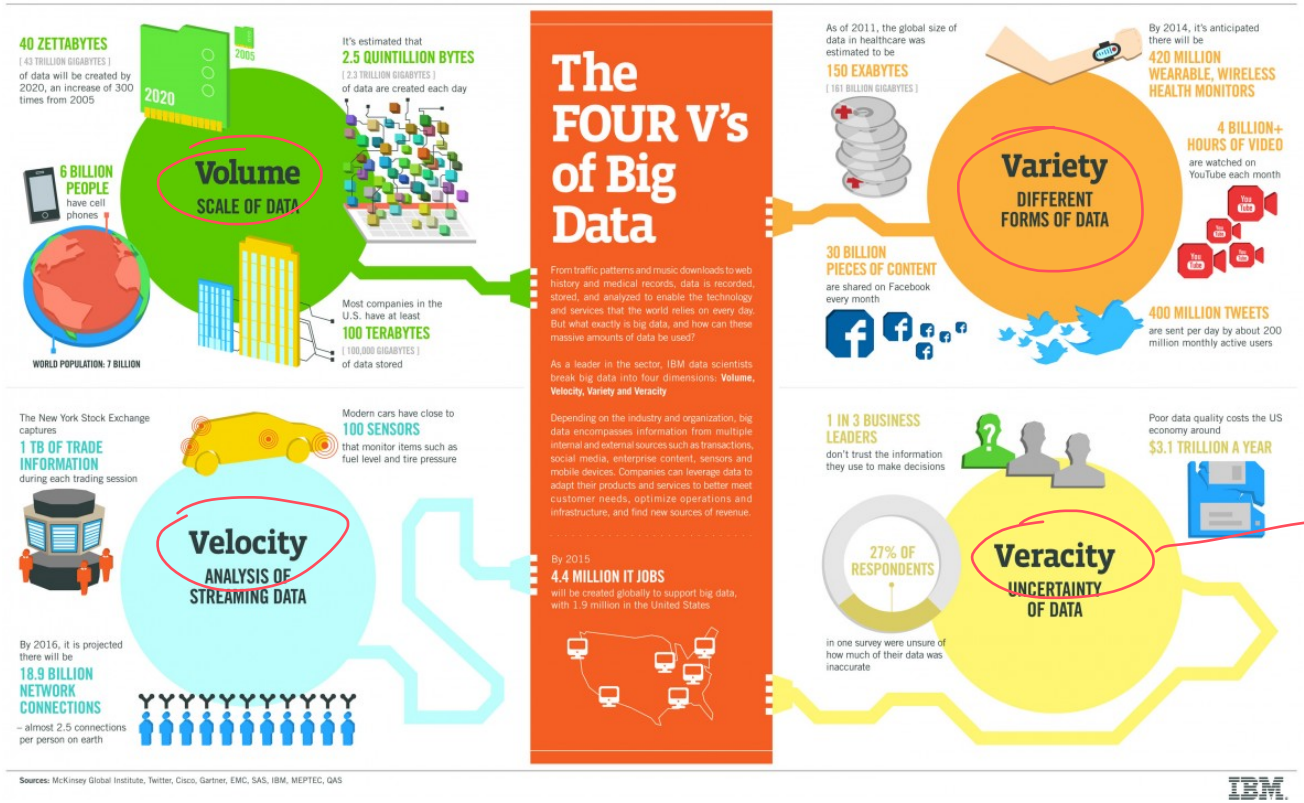
- Big data = data that you can't handle 'normally'

- Big Data usually requires learning of new tools
- Want to feel overwhelmed?

↳ Spark through pyspark API

What are Common Attributes of Big Data?

- Fifth **V** = Value



Variability

What are We Going to Consider?

- Basics of Big Data storage
  - ↳ databases, backups of data
- Basics of Big Data querying
  - ↳ SQL style commands
- Basics of handling streaming data
  - ↳ Spark for streaming
- Summarizing and modeling Big Data

Data and Model Pipelines

In the end, we should have a reasonable idea of how data comes in, gets transformed/combined/etc., can be used to build models or predict an outcome!

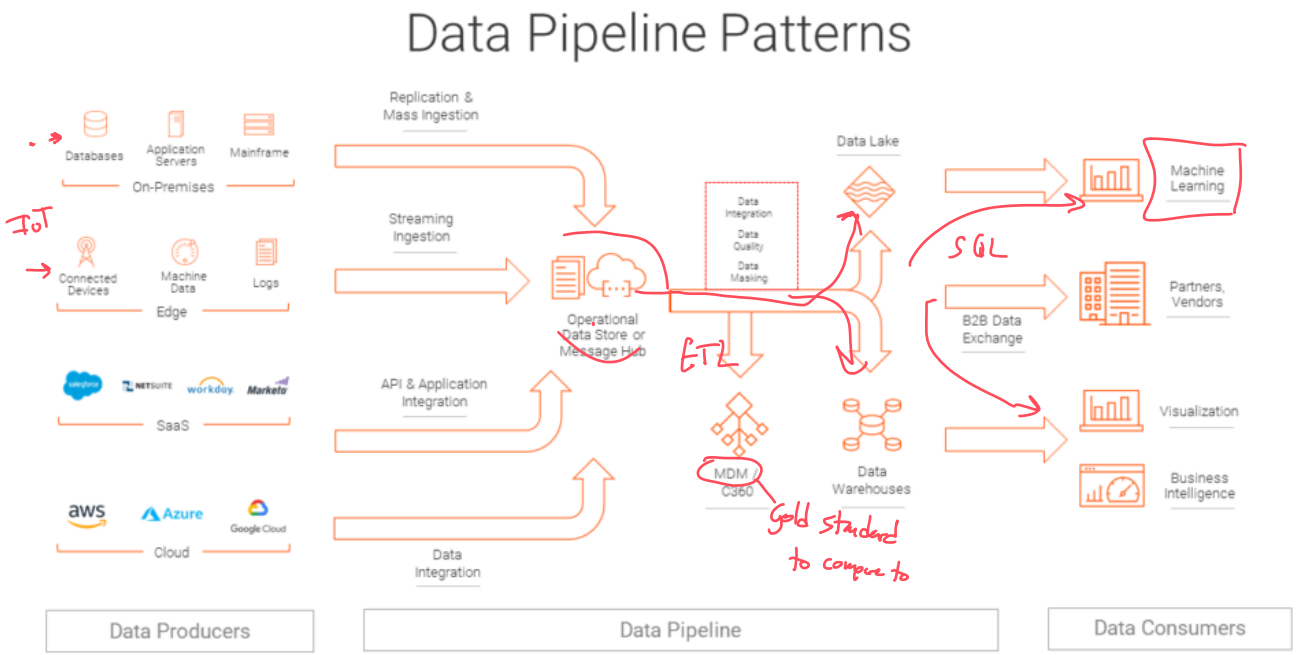


Image from informatica.com

Recap

- Big Data requires learning new tools & considering different algorithms
- Storage and retrieval of data is important
- Modeling and summarizing data can be done
- Should consider overall process/pipeline of data from start to finish