Microsoft

Roles and Responsibilities of the Azure Data Engineer

Jes Schultz

Jes Schultz

- Software Engineer
- Microsoft

  - jes.schultz@microsoft.com
    @grrl_geek
    LessThanDot.com

- Microsoft Certified, Azure Data Engineer Associate
- Microsoft Specialist, Design and Implement Cloud Data Platform Solutions
- Microsoft Certified Solutions Expert, Data Management and Analytics

- 6-time Microsoft Data Platform MVP
- Author, Pro SQL Server 2012 Practices

Abstract

The rise of data science has led to the rise of the data engineer.

While there's a huge push for people to become data engineers, there's also a lot of confusion about what this role is and does. If you're interested in taking your career in a new direction, come and learn what the data engineer role entails, the skills you need to learn, and the Azure services that tie to those skills. This will all be in alignment to the new Microsoft Azure Data Engineer Associate certification.

This session will put you on a path to transforming your career and becoming an Azure Data Engineer.

What you’re going to learn today

Topics

The problems a data engineer solves.
The skills a data engineer needs.
The tools a data engineer uses.
Microsoft's data engineer certification path.
The steps you can take towards becoming a data engineer.

Why data engineering?

Why data engineering?

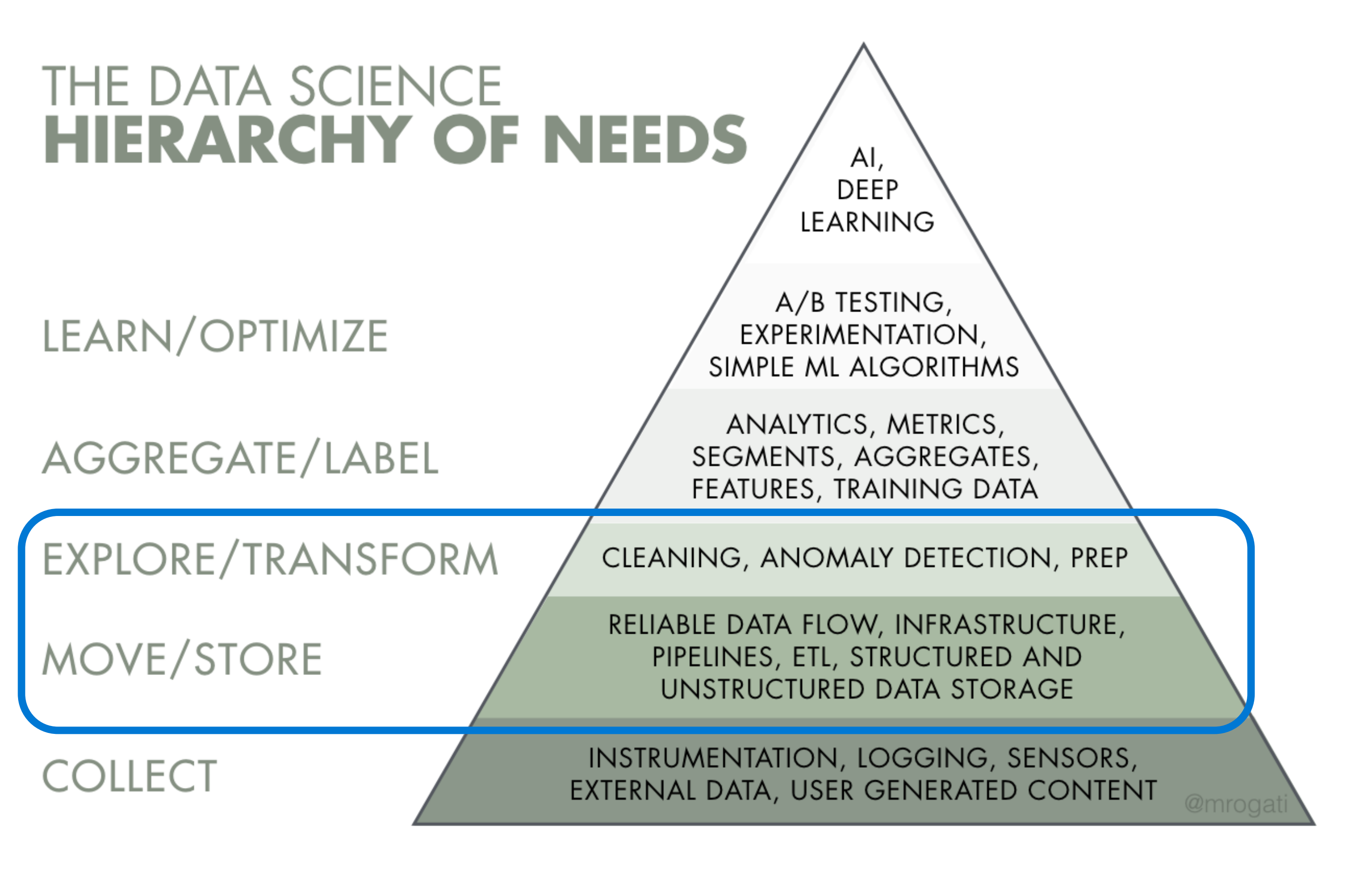
Behind every good data scientist is one or more data engineers!

It’s challenging!

Jobs typically offer a high salary!

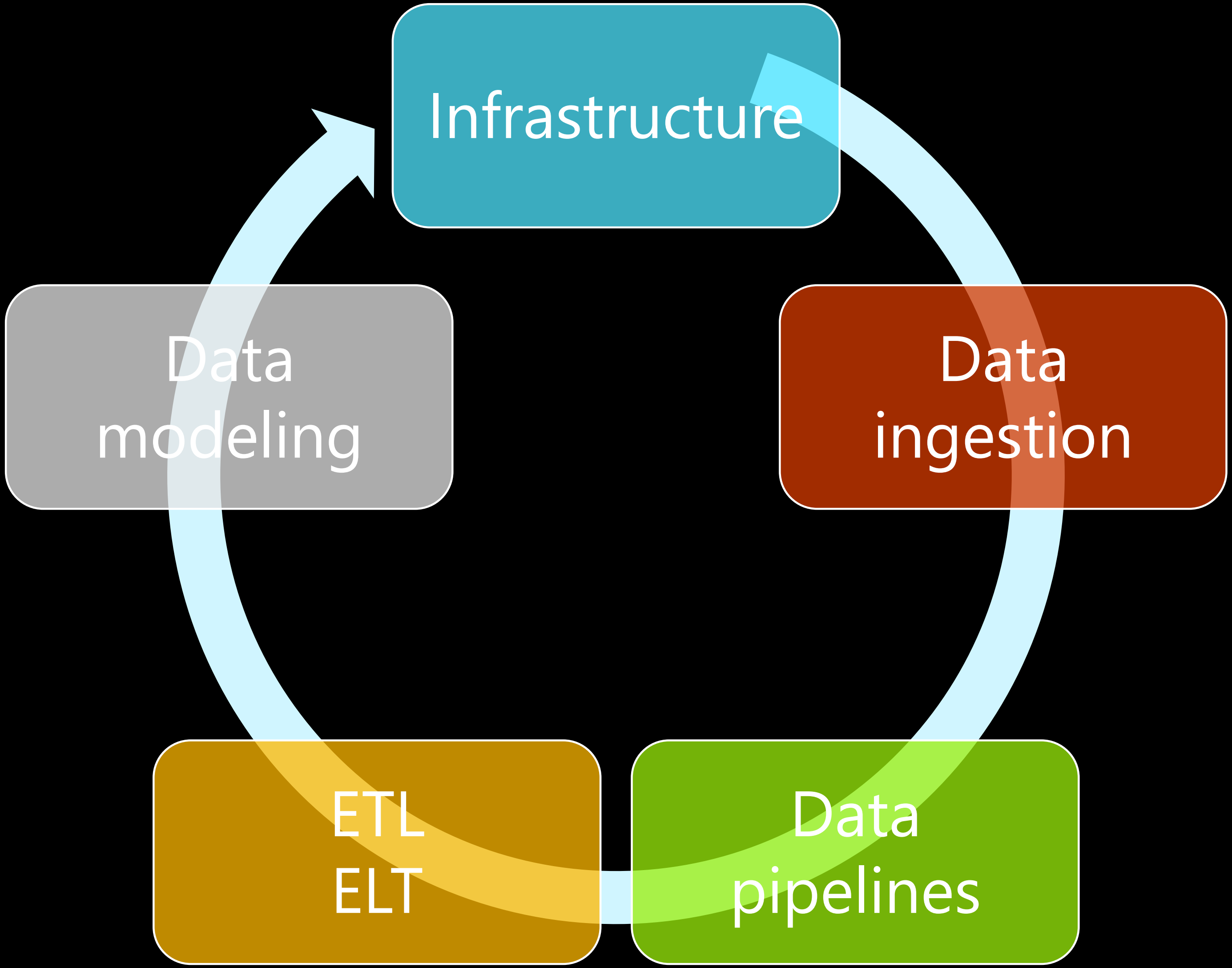
Problems a data engineer solves

Monica Rogati's
Data Science Hierarchy of Needs

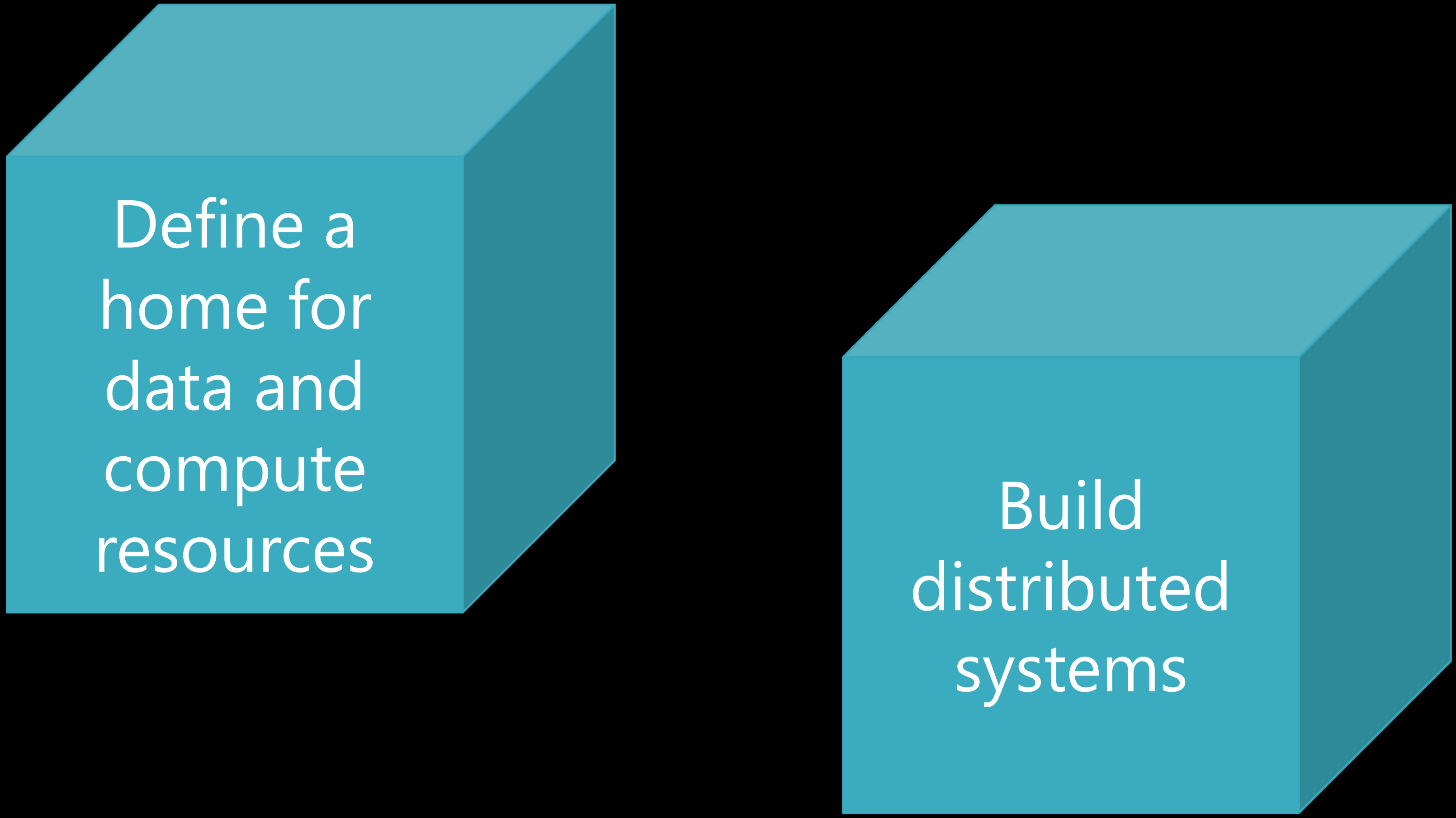

THE DATA SCIENCE
HIERARCHY OF NEEDS
AI, DEEP LEARNING
LEARN/OPTIMIZE
A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS
AGGREGATE/LABEL
ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA
EXPLORE/TRANSFORM
CLEANING, ANOMALY DETECTION, PREP
MOVE/STORE
RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE
COLLECT
INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT
@mrogati

https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007?source=post_page------------------------

Problems a data engineer solves



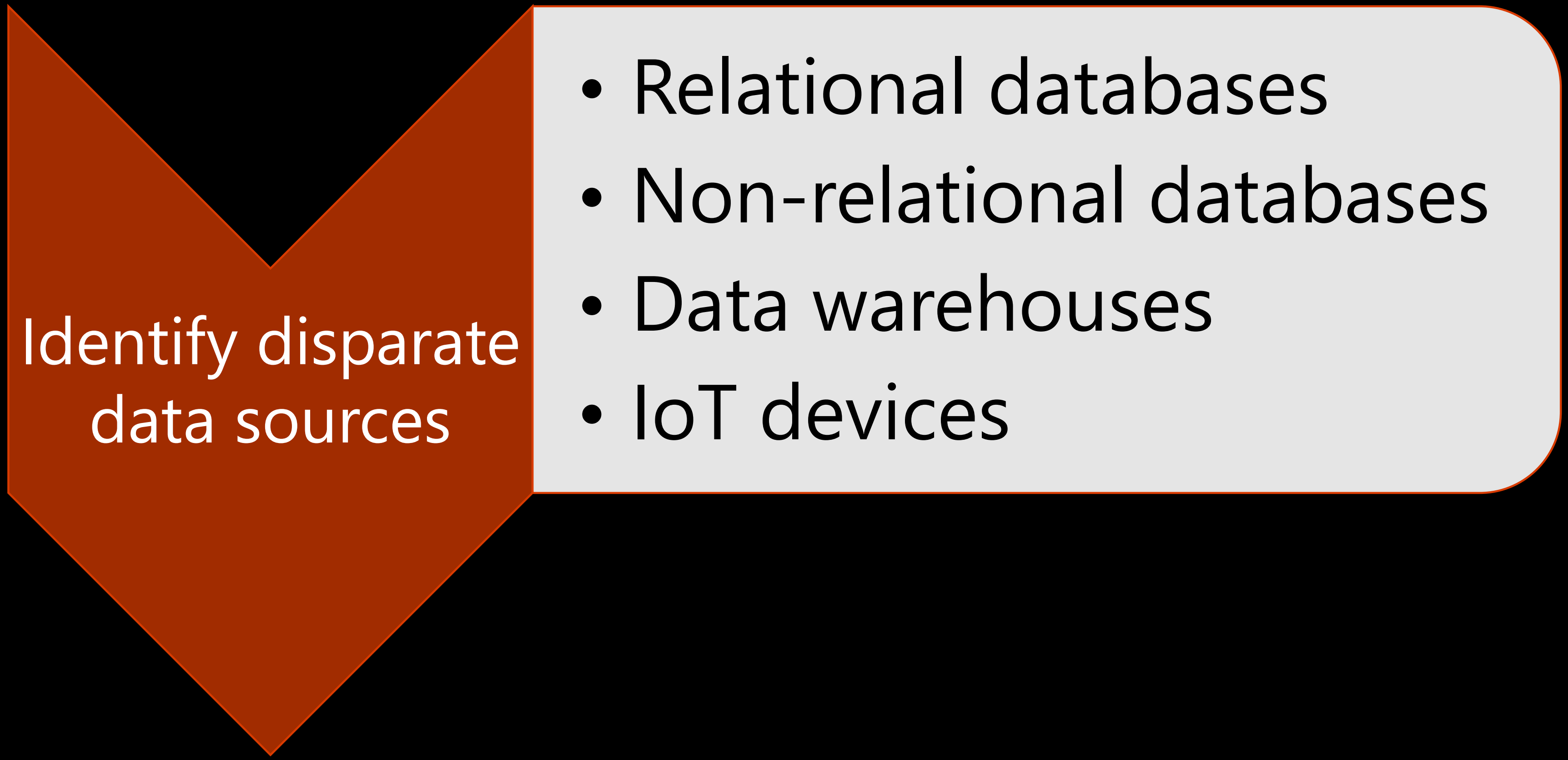
Infrastructure



Define a home for data and compute resources

Build distributed systems

Data ingestion



Identify disparate data sources

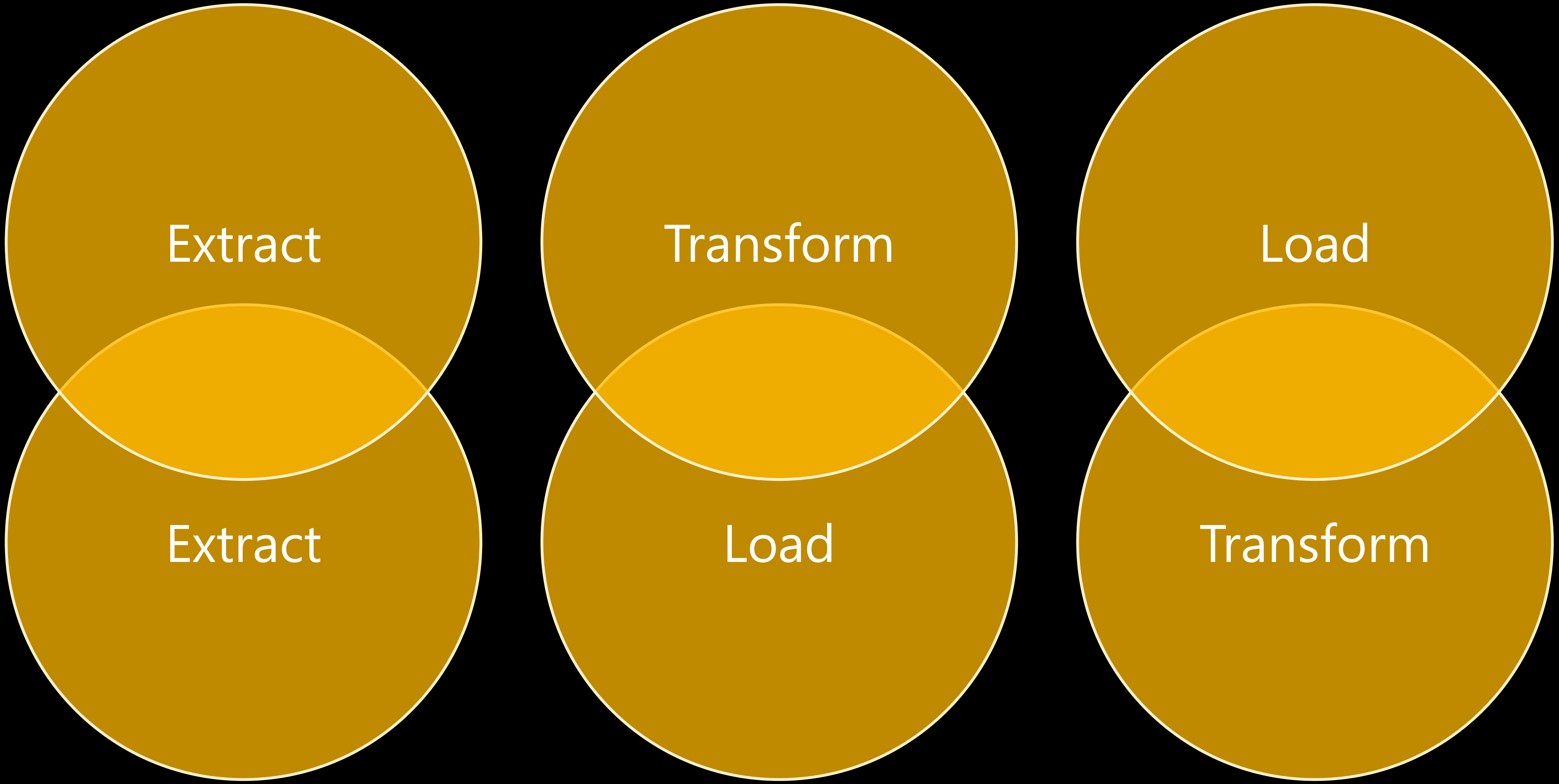
- Relational databases
- Non-relational databases
- Data warehouses
- IoT devices

Data pipelines

Build a pipeline to bring the data into a common data store

Source control
Build tools
Configuration management
Monitoring

ETL? ELT?



Extract

Extract

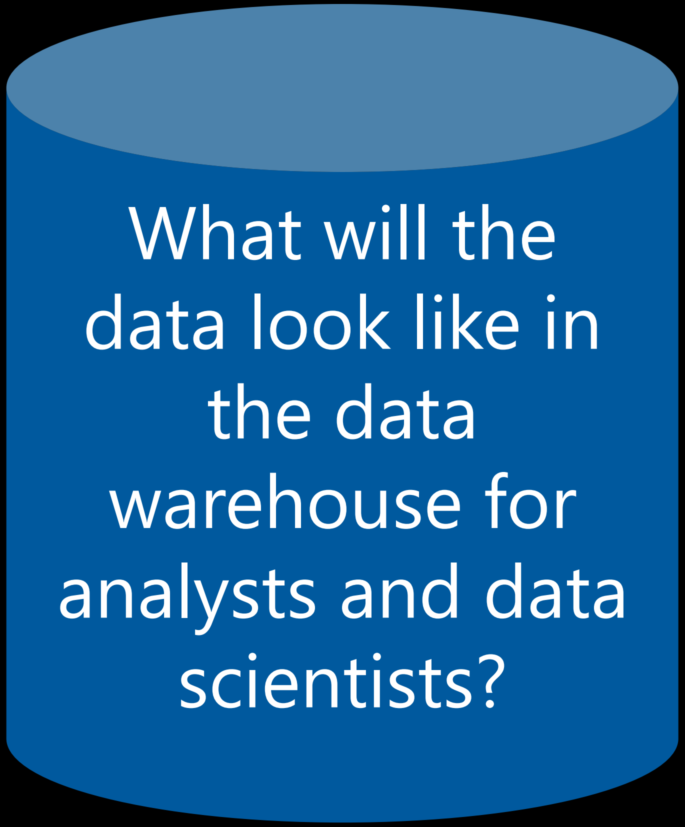
Transform

Load

Load

Transform

Data modeling



What will the data look like in the data warehouse for analysts and data scientists?

Skills a data engineer needs

An understanding of DevOps

- Every aspect of data engineering should involve a DevOps process
  - Source control
  - Continuous Integration
    - Testing
  - Continuous Deployment
    - Pushing out changes

Familiarity with scripting languages

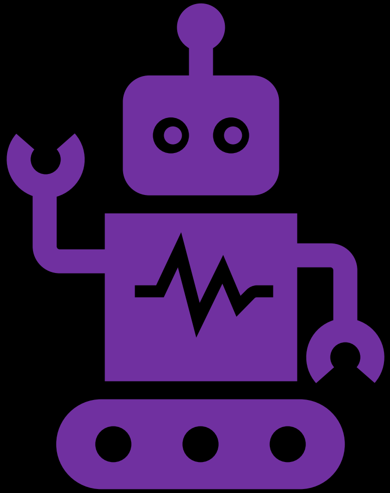
- Infrastructure as code
- Pipeline as code

- Time-saving
- Consistent
- Repeatable
- Verifiable

Knowledge of automation

- Processes should be designed once for use multiple times

- Time-saving
- Consistent
- Repeatable
- Verifiable



An ability to scale systems

- Vertical
  - Add more resources to a single server
- Horizontal
  - Add more servers

- No company has less data than they did a year ago.
- No company is interested in using less data to make decisions.
- No one can predict the future and how successful or business-critical a component will become.

Tools a data engineer uses

Infrastructure as code

- Azure Resource Manager (RM) templates
- AWS CloudFormation templates
- Google Cloud Deployment Manager templates
- Terraform
- Chef
- Puppet

Pipeline as code

- Azure DevOps
- Jenkins
- Travis CI
- TeamCity

Languages

- SQL
- Python 3
- R
- JavaScript
- Azure – PowerShell, CLI

Microsoft’s Data Engineer Certification path

Exam DP-200: Implementing an Azure Data Solution

- Implement data storage solutions
  - Implement non-relational data stores
  - Implement relational data stores
  - Manage data security
- Manage and develop data processing
  - Develop batch processing solutions
  - Develop streaming solutions
- Monitor and optimize data solutions
  - Monitor data storage
  - Monitor data processing
  - Optimize Azure data solutions

Exam DP-201: Designing an Azure Data Solution

- Design Azure data storage solutions
  - Recommend an Azure Data solution based on requirements
  - Design non-relational cloud data stores
  - Design relational cloud data stores
- Design data processing solutions
  - Design batch processing solutions
  - Design real-time processing solutions
- Design for data security and compliance
  - Design security for source data access
  - Design security for data policies and standards

Technologies covered

Databases

Cosmos DB

- NoSQL
- Document (JSON)
- Key-value
- Graph

SQL Database

- Relational
- Single / elastic pool
- Managed Instance
- Hyperscale

Synapse Analytics

- Multi-Parallel Processing (MPP)

Data stores

Blob storage

- Unstructured data
- Flat files
- Media files

Data Lake Storage

- Unstructured data
- HDFS
- Petabytes of storage

Data transformation

Data Factory

- ETL at scale
- 80+ connectors
- SSIS integration runtime

Data ingestion

Stream Analytics

- Serverless real-time analytics
- IoT devices
- Event Hubs

Data analytics

Databricks

- Big Data analytics platform
- Based on Apache Spark
- Notebooks

Your next steps

- Study for and take the Microsoft data engineer certification exams.
- Learn comparable tools from other vendors.
- Learn Python.
- Explore the parts you find most interesting.
- Understand this is a long-term process!

Resources

- Has the Data Engineer replaced the Business Intelligence Developer?
- The Rise of the Data Engineer
- A Beginner’s Guide to Data Engineering — Part I
- What is a Data Engineer?
- The AI Hierarchy of Needs
- 5 things you should know for a career in data engineering
- Team Data Science Process
- Python for Data Professionals

Questions?

Jes Schultz

- Software Engineer
- Microsoft

  - jes.schultz@microsoft.com
    @grrl_geek
    LessThanDot.com

- Microsoft Certified, Azure Data Engineer Associate
- Microsoft Specialist, Design and Implement Cloud Data Platform Solutions
- Microsoft Certified Solutions Expert, Data Management and Analytics

- 6-time Microsoft Data Platform MVP
- Author, Pro SQL Server 2012 Practices