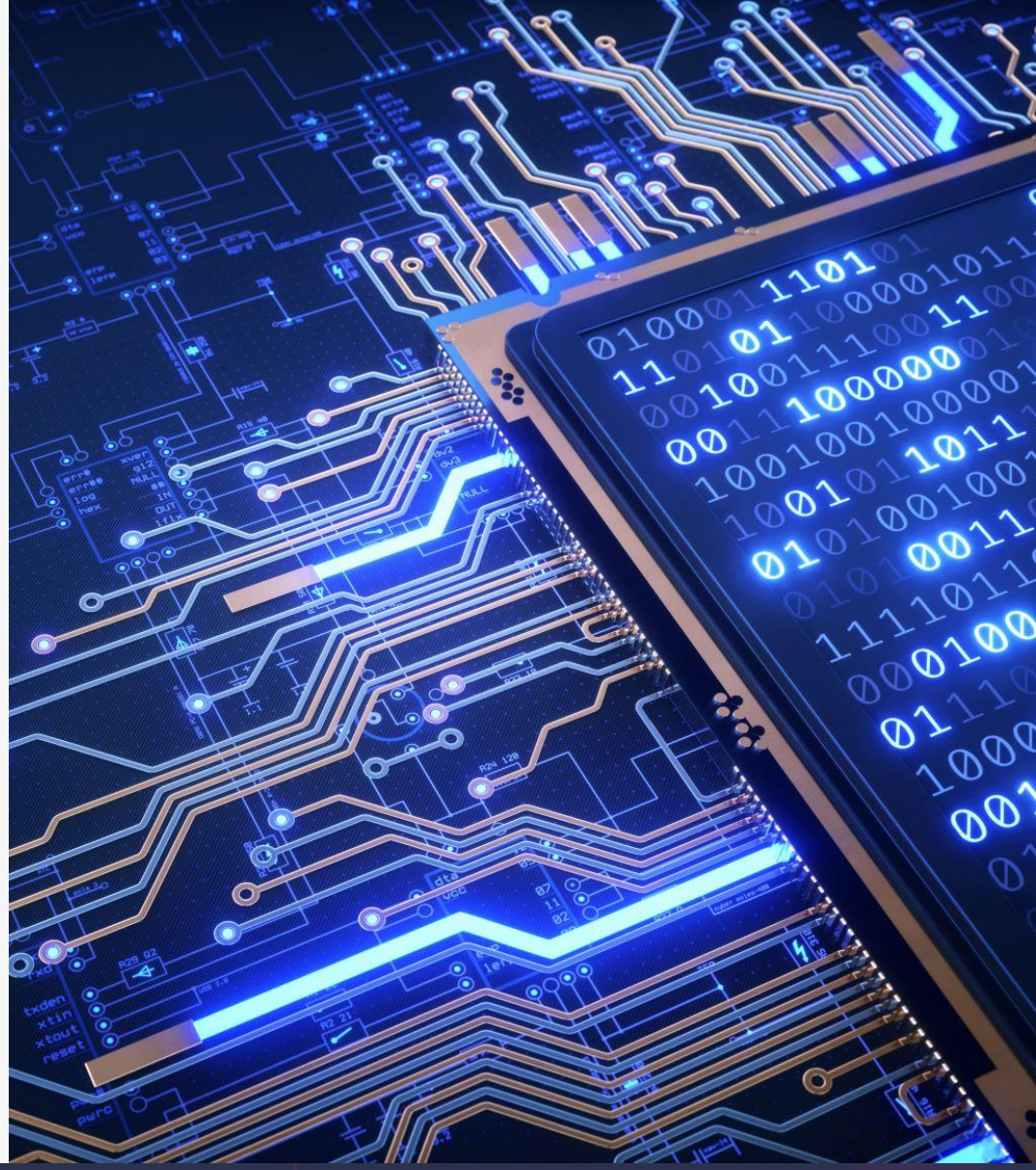
MSDS Data Engineering 610 Week 1

Kevin McBeth

Regis University



Agenda

- Admin
- Course Introduction
- Week 1 Content
- Week 1 Lab
- Week 1 Discussion (optional)

Course Introduction

Focus on Big Data Tools and Technologies, mostly Lab oriented.

- Week 1: Data Science vs Data Engineering, Lab Setup (Word Count)
- Week 2: Hadoop Ecosystem (Word Count, with some Python code)
- Week 3: SQL (SQL Gymnastics)
- Week 4: NoSQL (MongoDB)
- Week 5: API's and Data Architecture / Data Governance (API in Python)
- Week 6: Web Scraping & Spark Theory (Web Scraping in Python)
- Week 7 & 8: Spark (Pyspark Project)

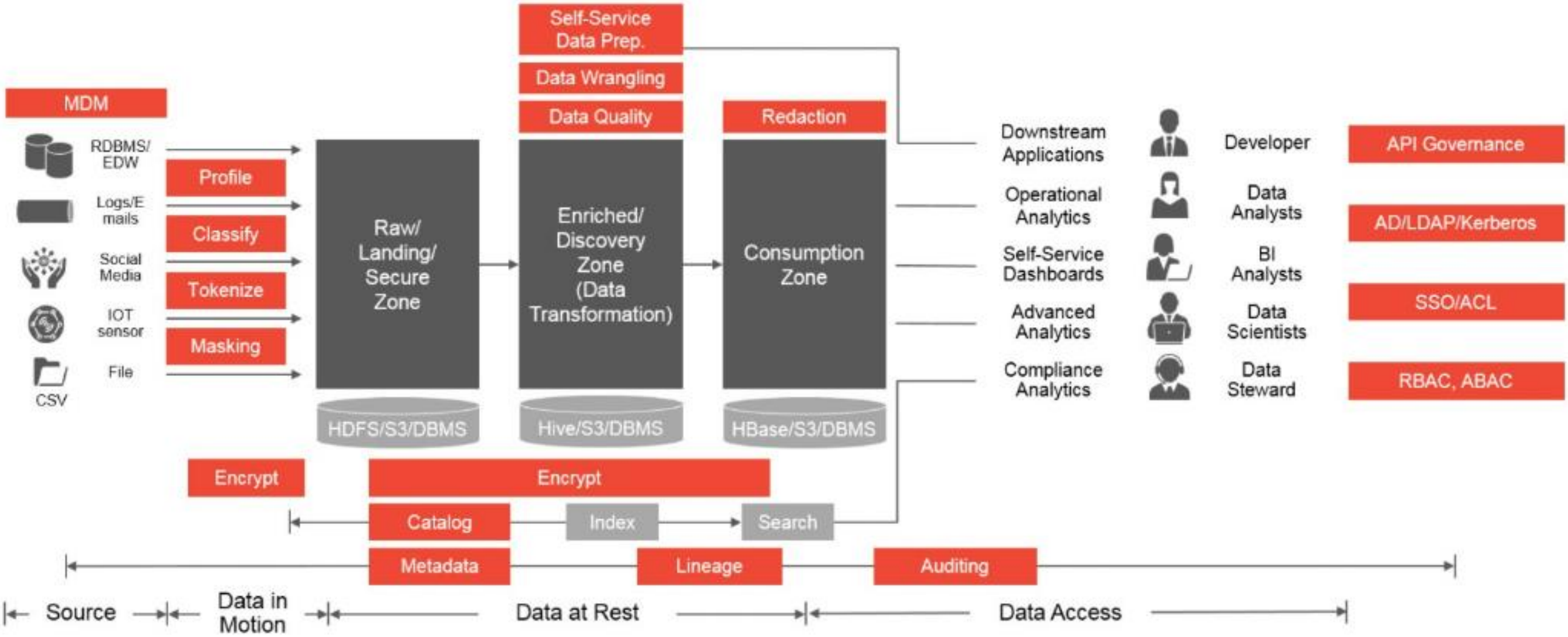
Admin

- Grading Policy
  - Very Flexible
  - Show Effort
  - If things aren't working communicate ahead of time (GCP issues are a known problem)
  - Drop the lowest assignment score (except for Spark Final Project)
  - Drop the lowest discussion score
- Discussion Due Dates – Original Post Wednesday 11:59 PM or in session discussion
- Lab Due Dates – Following Monday @ 11:59 PM
  - Please let me know ahead of time if you need an extension, communicate on Monday if there are issues. Holdiay weekends this will get pushed to Tuesday.
- Hard Deadline on Final Assignment – Sunday 11:59 PM

Week 1 Content

- Update to Stitch Report, top 10 skills
- Highlights from the readings
- Data Engineering Roadmap
- Robert Half Salary Guide
- Glassdoor Salary, Data Engineer vs Scientist vs Analyst

Unified Data Governance Reference Architecture



Gartner

Week 1 Lab

- Where is the Lab?
- What are the Deliverables?
  - UP TO two page report answering assignment questions
  - Screen shot of all work done in the lab. Must include the wordcount commands and output of **sort -gr -k 2 result | head**

Week 1 Lab Changes

#changes to lab @ exercise create directories

hadoop fs -mkdir /shakespeare hadoop fs -mkdir /shakespeare/input hadoop fs -copyFromLocal shakespeare.txt /shakespeare/input hadoop fs -ls /shakespeare/input

cd simple_Hadoop_MapReduce_example

NOTE: The slashes below indicate line continuation characters. Basically this is one command on multiple lines of parameters.

mapred streaming \

-mapper mapper.py \

-reducer reducer.py \

-input /shakespeare/input \

-output /shakespeare/output

Week 1 Discussion (optional)

- Data Science vs Data Engineer, explain the differences in your own words
- Past Experiences