A picture containing shape

Description automatically generated

C7084 Big Data

# SQL & Databricks material Topic outline

# SQL (freecodecamp)

⌨️ ([0:00](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=0s)) Introduction ⌨️ ([2:36](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=156s)) What is a Database? ⌨️ ([23:10](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=1390s)) Tables & Keys ⌨️ ([43:31](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=2611s)) SQL Basics ⌨️ ([52:26](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=3146s)) MySQL Windows Installation ⌨️ ([1:01:59](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=3719s)) MySQL Mac Installation ⌨️ ([1:15:49](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=4549s)) Creating Tables ⌨️ ([1:31:05](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=5465s)) Inserting Data ⌨️ ([1:38:17](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=5897s)) Constraints ⌨️ ([1:48:11](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=6491s)) Update & Delete ⌨️ ([1:56:11](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=6971s)) Basic Queries ⌨️ ([2:08:37](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=7717s)) Company Database Intro ⌨️ ([2:14:05](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=8045s)) Creating Company Database ⌨️ ([2:30:27](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=9027s) ) More Basic Queries ⌨️ ([2:26:24](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=8784s)) Functions ⌨️ ([2:45:13](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=9913s)) Wildcards ⌨️ ([2:53:53](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=10433s)) Union ⌨️ ([3:01:36](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=10896s)) Joins ⌨️ ([3:11:49](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=11509s)) Nested Queries ⌨️ ([3:21:52](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=12112s)) On Delete ⌨️ ([3:30:05](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=12605s)) Triggers ⌨️ ([3:42:12](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=13332s)) ER Diagrams Intro ⌨️ ([3:55:53](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=14153s)) Designing an ER Diagram ⌨️ ([4:08:34](https://www.youtube.com/watch?v=HXV3zeQKqGY&t=14914s)) Converting ER Diagrams to Schemas

# Databricks

## Mod 00 – Intro and Setup

* Cloud setup and files to load into your Lab environment (to be completed before class)

## Mod 01 – Spark Architecture

* + - * + What is Apache Spark?
        + Spark with Hadoop
        + Logging into Spark
        + Spark architecture - The Big Picture
        + Programming languages, Unified Stack, Cluster Manager, Data, BI tools
        + Resilient Distributed Datasets (RDD)
        + Spark cluster components
        + Driver, Spark Context, YARN, Worker nodes and Executors, Files
        + Glossary of Spark terminology

## Mod 02 – Getting started with RDDs

* Creating Resilient Distributed Datasets (RDD), Dataframes and Tables
* RDD Operations
* What are Actions?
* What are Transformations?
* Pair RDD
* Hands-on exercises using most common Actions and Transformations

## Mod 03 – Spark SQL (DDL)

* What is Spark SQL?
* DataFrames
* Tables
* TempViews
* The Catalyst Optimizer
* What is Hive?
* Creating DataFrames from five sources including:
* Structured files, parallelize(), textFile(), Hive and mySQL
* CREATE TABLE function
* CreateOrReplaceTempView function
* Creating User Defined Functions

## Mod 04 – SparkSQL (DML)

* + How to query using functions for both:
    - * DataFrames
      * Tables/View
  + Explain Plans
  + How to Write contents to a file

## Mod 05 – Complex Data Types

* + Learn how to use complex Data types including:
  + Standard functions for Collections
  + Array
  + Map
  + Structs

## Mod 06 – JSON (Bonus material)

* Querying JSON objects
* Commands include:
  + get\_json\_object
  + To\_json
  + From\_json
  + selectExpr
  + Nested JSON objects using:
  + Explode
  + getItem

## Mod 07 – Structured Streaming

* What is a Spark Structured Streaming ?
* Batch versus Streaming Datasets/DataFrames
* Structured Streaming Model
* Input: **read.stream()**function
* Output: **write.stream()**function and most common arguments
  + **format()**
  + **outputMode()**
* Demos or Hands-on labs including:
  + Sockets data source and HDFS data source
  + Windowing streaming and streaming using a script

## Mod 08 – Job Architecture and Spark UI

* The three main components of a Spark query
  + Job, Stage and Task
* Execution for these components across a distributed architecture
* Monitoring Spark jobs via the Spark UI including:
  + Jobs tab
  + Stages tab
  + Storage tab
  + Environment tab
  + Executors tab
  + SQL tab
  + JDBC/ODBC Server
  + Structured Streaming tab

## Mod 09 – Job Architecture and Spark UI

* Spark Catalog
  + List Hive tables and Spark Views
  + List column names on Table
  + List Spark functions
* Catalyst Optimizer functionalities including:
  + Optimize user code for more efficient processing
  + More performant join strategy (when using spark.read() function)
  + Dropping hints to Optimizer, Column pruning and Predicate pushdown
* Tungsten Project functionalities including:
  + Binary processing
  + Improved memory usage

Whole-stage code generation and Vectorization

## Mod 10 – Performance Tuning

* Caching and Persistence
* Checkpointing
* Accumulators
* Broadcast Variables
* Parallelism and Repartitioning using repartition() and coalesce()
* Shuffle Partitions
* Comparing Sequence, ORC and Parquet file format performance
* Java serialization versus Kryo serialization

## Mod 11 – Machine Learning (MLib and ML)

Machine Learning concepts

Terminology

MLib (RDD) vs ML (DataFrames)

Collaborative Filtering (ALS) example

ML Predictive concepts

Transformer, Estimator, Pipeline

ML (non-Predictive) examples

Correlation

KMeans

ML Predictive examples

Logistic Regression

Gradiant Boost Tree

Decision Tree