

Data Engineering Reference Links

- Nikhil Kumar Karnati

Contents

- Python1
- Scala2
- Spark2
- Hadoop3
- MS SQL.....3
- Basic Linux Commands.....3
- Data Warehousing.....4
- Azure Services Stack.....4

Python

S.no	Topic	Ref Links
1.	Python Installation	https://www.youtube.com/watch?v=YYXdXT2l-Gg https://www.guru99.com/how-to-install-python.html
2.	W3 Schools Tutorial	https://www.w3schools.com/python/default.asp
3.	Hacker Earth Tutorial	https://www.hackerearth.com/practice/python/getting-started/input-and-output/tutorial/
4.	PyTest	https://wiki.python.org/moin/PyTest
5.	Logging and Exception	https://docs.python.org/2/library/logging.html https://docs.python.org/3/tutorial/errors.html
6.	Regex	https://writersbyte.com/featured-post/regex-101-for-python-data-science/
7.	DSA	https://www.udemy.com/course/data-structures-algorithms-python/ https://www.udemy.com/course/leetcode-in-python-50-algorithms-coding-interview-questions/ https://www.youtube.com/watch?v=cWfE1F15kpw&t=712s https://github.com/VicodinAbuser/ZTM-DS-and-Algo-Python
8.	Python OOPS	https://www.analyticsvidhya.com/blog/2020/09/object-oriented-programming/

Scala

S.no	Topic	Link
		https://www.udemy.com/course/microsoft-azure-database-and-analytics/
1.	Scala Udemy	https://www.udemy.com/course/rock-the-jvm-scala-for-beginners/
2.	Scala doc	https://docs.scala-lang.org/scala3/book/introduction.html

Spark

S.no	Topic	Link
1.	Spark Tutorial (theory Concepts)	https://techvidvan.com/tutorials/spark-tutorial/
2.	RDD Operations (Transformations & Actions)	https://spark.apache.org/docs/latest/rdd-programming-guide.html
3.	Spark tuning	https://spark.apache.org/docs/latest/tuning.html#memory-management-overview
4.	Catalyst optimizer	https://blog.bi-geek.com/en/spark-sql-optimizador-catalyst/
5.	Window Functions	https://medium.com/expedia-group-tech/deep-dive-into-apache-spark-window-functions-7b4e39ad3c86
6.	Spark sql doc	Spark SQL — PySpark 3.1.1 documentation (apache.org)
7.	Skewness issue	https://blog.clairvoyantsoft.com/optimize-the-skew-in-spark-e523c6ee18ac
8.	Spark Out of Memory	https://labs.criteo.com/2018/01/spark-out-of-memory/ https://www.unraveldata.com/common-reasons-spark-applications-slow-fail-part-1/
9.	Spark join strategies	https://towardsdatascience.com/strategies-of-spark-join-c0e7b4572bcf
10	Spark Advanced	https://www.udemy.com/course/apache-spark-3-beyond-basics/
11	Delta lake- Databricks overview	https://docs.microsoft.com/en-us/azure/databricks/delta/delta-faq

Hadoop

S.no	Topic	Link
1.	Hadoop Tutorial (Hdfs, MapR, hive, sqoop)	https://www.javatpoint.com/hadoop-tutorial https://www.guru99.com/bigdata-tutorials.html
2.	Hive Scenario based ques	https://bigdataprogrammers.com/hive-most-asked-interview-questions-with-answers-part-ii/

MS SQL

S.no	Topic	Link
1.	SQL Basics	https://www.youtube.com/watch?v=zAOUpVM6R6I&list=PLtgiThe4j67rAoPmnCQmcgLS4ilc5ungg https://www.w3schools.com/sql/default.asp
2.	Practice questions	https://www.youtube.com/watch?v=fvPddKyHxpQ
3.	CTE in sql	https://www.geeksforgeeks.org/cte-in-sql/

Basic Linux Commands

S.no	Topic	Link
1.	Basic commands	https://www.hostinger.in/tutorials/linux-commands
		https://www.javatpoint.com/linux-commands
		https://www.geeksforgeeks.org/awk-command-unixlinux-examples/

Data Warehousing

S.no	Topic	Link
1.	Data Warehouse	https://www.tutorialspoint.com/dwh/dwh_quick_guide.htm
2.	DWH Interview Q	https://career.guru99.com/top-50-datawarehousing-questions-answers/
3.	Data Modelling	https://intellipaat.com/blog/data-modeling-tutorial-for-beginners/
4.	Normal forms	https://www.geeksforgeeks.org/normal-forms-in-dbms/

Azure Services Stack

S.no	Topic	Link
1.	Azure services Basics for DE	https://www.udemy.com/course/microsoft-azure-database-and-analytics/
2.	Microsoft Docs	https://docs.microsoft.com/en-us/learn/paths/azure-for-the-data-engineer/