DATA ENGINEERING CONCEPTS CHEATSHEET

1. DATA STORAGE

- Databases
- Relational (SQL): MySQL, PostgreSQL, Oracle, MS SQL Server
- NoSQL: MongoDB, Cassandra, Redis, DynamoDB
- Data Lakes
- Storage: AWS S3, Azure Blob Storage, Google Cloud Storage
- Frameworks: Hadoop HDFS

2. DATA WAREHOUSING

- Traditional
- SQL-based: Amazon Redshift, Google BigQuery, Snowflake, Teradata
- Modern
- Columnar Storage: Parquet, ORC
- Data Lakehouses: Delta Lake, Apache Hudi, Apache Iceberg

3. DATA INGESTION

- Batch Processing
- Tools: Apache Sqoop, Apache Nifi, AWS Glue
- Stream Processing
- Tools: Apache Kafka, Apache Flink, Apache Pulsar, AWS Kinesis

4. ETL (EXTRACT, TRANSFORM, LOAD) / ELT (EXTRACT, LOAD, TRANSFORM)

- ETL Tools
- Informatica, Talend, Apache Nifi, AWS Glue
- ELT Tools
- dbt (data build tool), Matillion, Airflow

5. DATA TRANSFORMATION

- Scripting Languages
- SQL, Python, R
- Frameworks
- Apache Spark, Apache Beam

6. DATA ORCHESTRATION

- Tools: Apache Airflow, Luigi, Prefect, Dagster

7. DATA PIPELINES

- Design Patterns: Lambda Architecture, Kappa Architecture
- Tools: Apache Beam, Google Dataflow

8. DATA GOVERNANCE & QUALITY

- Data Governance
- Tools: Apache Atlas, AWS Lake Formation
- Data Quality
- Tools: Great Expectations, Deequ, Soda SQL

9. DATA INTEGRATION

- APIs
- RESTful APIs, GraphQL
- Middleware
- ESB (Enterprise Service Bus): MuleSoft, Apache Camel

10. DATA SECURITY

- Encryption: SSL/TLS, AES, RSA
- Access Control: IAM (Identity and Access Management), Role-Based Access Control (RBAC)

11. DATA MONITORING & LOGGING

- Tools: Prometheus, Grafana, ELK Stack (Elasticsearch, Logstash, Kibana), Splunk

12. BIG DATA PROCESSING FRAMEWORKS

- Hadoop Ecosystem: Hadoop MapReduce, Hive, Pig
- Apache Spark

13. CLOUD SERVICES

- AWS: S3, RDS, Redshift, Glue, EMR, Kinesis
- Azure: Blob Storage, SQL Database, Synapse Analytics, Data Factory, HDInsight
- Google Cloud: Cloud Storage, BigQuery, Dataflow, Pub/Sub, Dataproc

14. DATA FORMATS

- Structured: CSV, JSON, XML
- Semi-Structured: Avro, Parquet, ORC

15. DATA VERSIONING

- Tools: DVC (Data Version Control), Delta Lake

16. CI/CD FOR DATA ENGINEERING

- Tools: Jenkins, GitLab CI/CD, CircleCI, Azure DevOps

17. DATA VISUALIZATION

- Tools: Tableau, Power BI, Looker, Grafana, Superset

18. MACHINE LEARNING INTEGRATION

- Platforms: MLflow, TensorFlow Extended (TFX), Kubeflow

19. DATA LINEAGE

- Tools: OpenLineage, DataHub, Amundsen

20. POPULAR PROGRAMMING LANGUAGES

- Python, SQL, Java, Scala, Go