DP-203 Resources

1. **Design and Implement Data Storage (40-45%)**
    1. Design a data storage structure
        1. design an Azure Data Lake solution
            1. https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices
            2. https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-data-scenarios
        2. recommend file types for storage &
        3. recommend file types for analytical queries
            1. https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage#dataset-properties
        4. design for efficient querying
            1. https://docs.microsoft.com/en-us/azure/data-explorer/data-lake-query-data#optimize-your-query-performance
            2. https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-query-acceleration
            3. https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-query-acceleration-how-to?tabs=azure-powershell%2Cpowershell
        5. design for data pruning
            1. https://en.wikipedia.org/wiki/Decision_tree_pruning
            2. https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-performance-tuning-guidance
            3. https://docs.microsoft.com/bs-cyrl-ba/azure/databricks//delta/optimizations/dynamic-file-pruning
            4. https://databricks.com/blog/2020/04/30/faster-sql-queries-on-delta-lake-with-dynamic-file-pruning.html
            5. https://docs.microsoft.com/en-ca/azure/databricks//delta/optimizations/dynamic-file-pruning
        6. design a folder structure that represents the levels of data transformation
            1. https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices#directory-layout-considerations
            2. https://techcommunity.microsoft.com/t5/data-architecture-blog/how-to-organize-your-data-lake/ba-p/1182562
            3. https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-namespace
        7. design a distribution strategy
            1. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute
        8. design a data archiving solution

        1. https://azure.microsoft.com/en-ca/updates/archive-tier-for-azure-data-lake-storage-now-generally-available/
        2. https://docs.microsoft.com/en-us/azure/storage/blobs/storage-blob-storage-tiers?tabs=azure-portal#archive-access-tier
2. Design a partition strategy
    1. design a partition strategy for files
    2. design a partition strategy for analytical workloads
    3. design a partition strategy for efficiency/performance
    4. design a partition strategy for Azure Synapse Analytics
    5. identify when partitioning is needed in Azure Data Lake Storage Gen2
        1. https://docs.microsoft.com/en-us/azure/architecture/best-practices/data-partitioning
        2. https://docs.microsoft.com/en-us/azure/architecture/best-practices/data-partitioning-strategies
3. Design the serving layer
    1. design star schemas
        1. https://docs.microsoft.com/en-us/power-bi/guidance/star-schema
        2. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-overview
    2. design slowly changing dimensions
        1. https://en.wikipedia.org/wiki/Slowly_changing_dimension
        2. https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/
        3. https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types
        4. https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/2-describe
        5. https://www.youtube.com/watch?v=Sg2AAk1vwEs
    3. design a dimensional hierarchy
        1. https://docs.microsoft.com/en-us/power-bi/guidance/star-schema#snowflake-dimensions
        2. https://en.wikipedia.org/wiki/Snowflake_schema
        3. https://docs.microsoft.com/en-us/azure/data-factory/connector-snowflake
    4. design a solution for temporal data
        1. https://docs.microsoft.com/en-us/azure/azure-sql/temporal-tables
        2. https://en.wikipedia.org/wiki/Temporal_database
    5. design for incremental loading
        1. https://docs.microsoft.com/en-us/azure/data-factory/tutorial-incremental-copy-overview
        2. https://docs.microsoft.com/en-us/azure/data-factory/tutorial-incremental-copy-change-tracking-feature-portal

        3. https://docs.microsoft.com/en-us/azure/data-factory/tutorial-incremental-copy-portal
        4. https://www.youtube.com/watch?v=F9cBFnxaSGI
    6. design analytical stores
        1. https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/analytical-data-stores
        2. https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/#lambda-architecture
    7. design metastores in Azure Synapse Analytics and Azure Databricks
        1. https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-use-external-metadata-stores
        2. https://docs.microsoft.com/en-us/azure/databricks/data/metastore/
        3. https://docs.microsoft.com/en-us/azure/synapse-analytics/metadata/overview
        4. https://docs.microsoft.com/en-us/azure/databricks/data/metastores/external-hive-metastore
        5. https://www.youtube.com/watch?v=pBB5zFnhgyE&list=PL7_h0bRfL52oZqAfV_kumYLUH5dbcWm9q
4. Implement physical data storage structures
    1. implement compression
        1. https://docs.microsoft.com/en-us/azure/data-factory/supported-file-formats-and-compression-codecs
        2. https://docs.microsoft.com/en-us/azure/data-factory/format-parquet
        3. https://databricks.com/glossary/what-is-parquet
        4. https://docs.informatica.com/data-integration/powerexchange-adapters-for-informatica/10-5/powerexchange-for-microsoft-azure-blob-storage-user-guide/microsoft-azure-blob-storage-data-objects/data-compression-in-microsoft-azure-blob-storage-sources-and-tar.html
    2. implement partitioning
        1. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition
    3. implement sharding
        1. https://docs.microsoft.com/en-us/azure/architecture/patterns/sharding
        2. https://docs.microsoft.com/en-us/azure/azure-sql/database/elastic-scale-introduction
        3. https://docs.microsoft.com/en-us/azure/azure-sql/database/elastic-scale-shard-map-management
    4. implement different table geometries with Azure Synapse Analytics pools
        1. https://docs.microsoft.com/en-us/azure/synapse-analytics/get-started-analyze-sql-pool

        2. https://docs.microsoft.com/en-us/azure/synapse-analytics/get-started-analyze-sql-on-demand
        3. https://docs.microsoft.com/en-us/azure/synapse-analytics/get-started-analyze-spark
    5. implement data redundancy
        1. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/backup-and-restore
        2. https://docs.microsoft.com/en-us/azure/cloud-adoption-framework/migrate/azure-best-practices/analytics/azure-synapse
        3. https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy
        4. https://docs.microsoft.com/en-us/azure/databricks/scenarios/howto-regional-disaster-recovery
    6. implement distributions
        1. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute
    7. implement data archiving
        1. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/backup-and-restore
        2. https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-supported-blob-storage-features
            a. https://docs.microsoft.com/en-us/azure/storage/blobs/storage-blob-storage-tiers
5. Implement logical data structures
    1. build a temporal data solution
        1. https://docs.microsoft.com/en-us/azure/azure-sql/temporal-tables
        2. https://docs.microsoft.com/en-us/azure/architecture/
    2. build external tables
        1. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop
    3. implement file and folder structures for efficient querying and data pruning
        1. https://docs.microsoft.com/en-us/azure/data-explorer/data-lake-query-data
        2. https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-performance-tuning-guidance
6. Implement the serving layer
    1. deliver data in a relational star schema
        1. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-overview
        2. https://en.wikipedia.org/wiki/Star_schema
    2. deliver data in Parquet files
        1. https://databricks.com/glossary/what-is-parquet
        2. https://docs.microsoft.com/en-us/azure/data-factory/format-parquet

      3. implement a dimensional hierarchy
         1. https://docs.microsoft.com/en-us/power-bi/guidance/star-schema#snowflake-dimensions
         2. https://en.wikipedia.org/wiki/Snowflake_schema
         3. https://docs.microsoft.com/en-us/azure/data-factory/connector-snowflake
2. **Design and Develop Data Processing (25-30%)**
   1. Ingest and transform data
      1. transform data by using Apache Spark
         1. https://docs.microsoft.com/en-us/azure/databricks/scenarios/databricks-extract-load-sql-data-warehouse
      2. transform data by using Transact-SQL
         1. https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-sql-data-warehouse
      3. transform data by using Data Factory
         1. https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-sql-database
         2. https://docs.microsoft.com/en-us/azure/data-factory/transform-data-using-spark
      4. transform data by using Azure Synapse Pipelines
         1. https://docs.microsoft.com/en-us/azure/synapse-analytics/get-started-pipelines
         2. https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities?toc=/azure/synapse-analytics/toc.json&bc=/azure/synapse-analytics/breadcrumb/toc.json
      5. transform data by using Stream Analytics
         1. https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-introduction
      6. cleanse data
         1. https://en.wikipedia.org/wiki/Data_cleansing
         2. https://www.sqlshack.com/data-cleansing-in-azure-machine-learning/
         3. https://app.pluralsight.com/guides/cleaning-data-with-azure-ml-studio
         4. https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/clean-missing-data
      7. split data
         1. https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/split-data
      8. shred JSON
         1. https://docs.microsoft.com/en-us/sql/relational-databases/json/convert-json-data-to-rows-and-columns-with-openjson-sql-server?view=sql-server-ver15

        2. https://docs.microsoft.com/en-us/sql/t-sql/functions/openjson-transact-sql?view=sql-server-ver15
    9. encode and decode data
        1. https://docs.microsoft.com/en-us/answers/questions/129474/azure-data-factory-base64-encoded-secrets.html
    10. configure error handling for the transformation
        1. https://docs.microsoft.com/en-us/azure/data-factory/how-to-data-flow-error-rows
        2. https://techcommunity.microsoft.com/t5/azure-data-factory/understanding-pipeline-failures-and-error-handling/ba-p/1630459
        3. https://docs.microsoft.com/en-us/azure/data-factory/data-factory-ux-troubleshoot-guide
        4. https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor
    11. normalize and denormalize values
        1. https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/normalize-data
    12. transform data by using Scala
        1. https://docs.microsoft.com/en-us/azure/databricks/scenarios/databricks-extract-load-sql-data-warehouse
    13. perform data exploratory analysis
        1. https://azure.microsoft.com/en-us/resources/videos/perform-exploratory-analytics-over-your-data-lake/
        2. https://docs.microsoft.com/en-us/learn/modules/perform-machine-learning-with-azure-databricks/
2. Design and develop a batch processing solution
    1. develop batch processing solutions by using Data Factory, Data Lake, Spark, Azure
        1. https://docs.microsoft.com/en-us/azure/data-factory/v1/data-factory-data-processing-using-batch
        2. https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/batch-processing
    2. Synapse Pipelines, PolyBase, and Azure Databricks &
    3. create data pipelines
        1. https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-versioned-feature-summary?view=sql-server-ver15
        2. https://docs.microsoft.com/en-us/azure/databricks/clusters/configure
        3. https://www.youtube.com/watch?v=JUQXx0R0RfE
    4. design and implement incremental data loads
        1. https://docs.microsoft.com/en-us/azure/data-factory/tutorial-incremental-copy-overview
    5. design and develop slowly changing dimensions

1. https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/
2. https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types
3. https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/2-describe

6. handle security and compliance requirements
    1. https://azure.microsoft.com/en-ca/overview/trusted-cloud/compliance/
    2. https://docs.microsoft.com/en-ca/azure/compliance/
7. scale resources
    1. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/quickstart-scale-compute-portal
    2. https://docs.microsoft.com/en-us/azure/data-factory/copy-activity-performance
8. configure the batch size
    1. https://docs.microsoft.com/en-us/azure/batch/batch-automatic-scaling
    2. https://docs.microsoft.com/en-us/azure/databricks/delta/delta-batch
9. design and create tests for data pipelines
    1. https://docs.microsoft.com/en-us/azure/databricks/dev-tools/ci-cd/ci-cd-azure-devops
10. integrate Jupyter/IPython notebooks into a data pipeline
    1. https://docs.microsoft.com/en-us/azure/databricks/notebooks/
    2. https://docs.microsoft.com/en-us/azure/databricks/notebooks/notebooks-use
    3. https://docs.microsoft.com/en-us/azure/databricks/notebooks/notebooks-manage
11. handle duplicate data
    1. https://docs.microsoft.com/en-us/azure/data-factory/how-to-data-flow-dedupe-nulls-snippets
12. handle missing data &
13. handle late-arriving data
    1. https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-time-handling
    2. https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-solution-patterns
    3. https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-module-reference/clean-missing-data
    4. https://learning.oreilly.com/library/view/stream-analytics-with/9781788395908/0b61b6d7-d805-42e2-a1cf-24148ce07f47.xhtml

        5. https://docs.microsoft.com/en-us/azure/stream-analytics/event-ordering
    14. upsert data
        1. https://docs.microsoft.com/en-us/azure/data-factory/data-flow-alter-row
    15. regress to a previous state
        1. https://docs.microsoft.com/en-us/answers/questions/31313/transactions-in-adf.html
        2. https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-sql-data-warehouse
    16. design and configure exception handling
        1. https://docs.microsoft.com/en-us/azure/data-factory/how-to-data-flow-error-rows
    17. configure batch retention
        1. Configure a simple Azure Batch Job with Azure Data Factory - Microsoft Tech Community
    18. design a batch processing solution
        1. https://docs.microsoft.com/en-us/azure/data-factory/v1/data-factory-data-processing-using-batch
    19. debug Spark jobs by using the Spark UI
        1. https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-job-debugging
3. **Design and develop a stream processing solution**
    1. develop a stream processing solution by using Stream Analytics, Azure Databricks, and Azure Event Hubs
        1. https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-introduction
        2. https://docs.microsoft.com/en-us/azure/databricks/spark/latest/structured-streaming/
        3. https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/data/stream-processing-databricks
    2. process data by using Spark structured streaming
        1. https://docs.microsoft.com/en-us/azure/databricks/spark/latest/structured-streaming/
    3. monitor for performance and functional regressions
        1. https://docs.microsoft.com/en-us/azure/databricks/kb/jobs/job-run-dash
        2. https://docs.microsoft.com/en-us/azure/data-factory/concepts-data-flow-monitoring
    4. design and create windowed aggregates
        1. https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions
    5. handle schema drift

1. https://docs.microsoft.com/en-us/azure/data-factory/concepts-data-flow-schema-drift

6. process time series data
   1. https://azure-samples.github.io/azureiotlabs/timeseriesinsights/#:~:text=Azure%20Time%20Series%20Insights%20is,over%20the%20world%20in%20seconds.
   2. https://docs.microsoft.com/en-ca/azure/time-series-insights/
7. process within one partition
8. process across partitions
   1. https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/event-hubs/partitioning-in-event-hubs-and-kafka
   2. https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features#partitions
   3. https://docs.microsoft.com/en-us/azure/stream-analytics/repartition
   4. https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization
9. configure checkpoints/watermarking during processing
   1. https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-time-handling
10. scale resources
    1. https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-scale-jobs
11. handle interruptions
    1. https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-job-reliability
    2. https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-time-handling
12. design and configure exception handling
    1. https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-output-error-policy
    2. https://docs.microsoft.com/en-us/azure/stream-analytics/configuration-error-codes
13. upsert data
    1. https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-documentdb-output
14. replay archived stream data
    1. https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-concepts-checkpoint-replay
15. design a stream processing solution
    1. https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/data/stream-processing-stream-analytics

4. Manage batches and pipelines
   1. trigger batches
   2. handle failed batch loads

            1. https://docs.microsoft.com/en-us/azure/batch/error-handling
            2. https://docs.microsoft.com/en-us/azure/batch/batch-job-task-error-checking
            3. https://docs.microsoft.com/en-us/azure/batch/batch-pool-node-error-checking
            4. https://docs.microsoft.com/en-us/azure/batch/best-practices
        3. validate batch loads
            1. https://docs.microsoft.com/en-us/azure/batch/batch-job-task-error-checking
        4. manage data pipelines in Data Factory/Synapse Pipelines
        5. schedule data pipelines in Data Factory/Synapse Pipelines
            1. https://docs.microsoft.com/en-us/azure/synapse-analytics/get-started-pipelines
            2. https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities
        6. implement version control for pipeline artifacts
            1. https://docs.microsoft.com/en-us/azure/data-factory/source-control
        7. manage Spark jobs in a pipeline
            1. https://docs.microsoft.com/en-us/azure/data-factory/v1/data-factory-spark

3. **Design and Implement Data Security (10-15%)**
    1. Design security for data policies and standards
        1. design data encryption for data at rest and in transit
            1. https://docs.microsoft.com/en-us/azure/storage/common/storage-service-encryption
            2. https://docs.microsoft.com/en-us/azure/cosmos-db/database-encryption-at-rest
            3. https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption
            4. https://docs.microsoft.com/en-us/azure/security/fundamentals/encryption-atrest
        2. design a data auditing strategy
            1. https://docs.microsoft.com/en-us/azure/azure-sql/database/auditing-overview
            2. https://docs.microsoft.com/en-us/azure/cosmos-db/audit-control-plane-logs
        3. design a data masking strategy, design for data privacy
            1. https://docs.microsoft.com/en-us/azure/security/fundamentals/protection-customer-data
            2. https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview
        4. design a data retention policy
            1. https://docs.microsoft.com/en-us/azure/storage/blobs/storage-lifecycle-management-concepts?tabs=azure-portal

      2. https://docs.microsoft.com/en-us/azure/azure-monitor/logs/manage-cost-storage
      3. https://docs.microsoft.com/en-us/azure/azure-monitor/app/data-retention-privacy
      4. https://azure.microsoft.com/en-ca/updates/retention-by-type/
   5. design to purge data based on business requirements
      1. https://docs.microsoft.com/en-us/azure/storage/blobs/soft-delete-blob-overview
      2. https://docs.microsoft.com/en-us/rest/api/keyvault/purgedeletedstorageaccount/purgedeletedstorageaccount
      3. https://docs.microsoft.com/en-us/azure/data-explorer/kusto/concepts/data-purge
      4. https://docs.microsoft.com/en-us/azure/storage/blobs/soft-delete-blob-enable
   6. design Azure role-based access control (Azure RBAC) and POSIX-like Access Control List
      1. https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control-model
   7. (ACL) for Data Lake Storage Gen2
      1. https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control
   8. Design and implement row-level and column-level security
      1. https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver15
      2. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/column-level-security
2. Implement data security
   1. implement data masking
      1. https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview
   2. implement Azure RBAC
      1. https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control-model
   3. implement POSIX-like ACLs for Data Lake Storage Gen2
      1. https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control
   4. implement a data retention policy
      1. https://azure.microsoft.com/en-ca/updates/lifecycle-management-for-azure-data-lake-storage-is-now-generally-available/
      2. https://docs.microsoft.com/en-us/azure/storage/blobs/storage-lifecycle-management-concepts?tabs=azure-portal
   5. implement a data auditing strategy
      1. https://docs.microsoft.com/en-us/azure/data-lake-analytics/data-lake-analytics-diagnostic-logs

6. manage identities, keys, and secrets across different data platform technologies
    1. https://docs.microsoft.com/en-us/rest/api/storageservices/authorize-with-shared-key
    2. https://docs.microsoft.com/en-us/azure/storage/common/storage-sas-overview?toc=/azure/storage/blobs/toc.json
    3. https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control-model
7. implement secure endpoints (private and public)
    1. https://docs.microsoft.com/en-us/azure/private-link/private-endpoint-overview
    2. https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices
    3. https://docs.microsoft.com/en-us/azure/data-factory/data-movement-security-considerations
8. implement resource tokens in Azure Databricks
    1. https://docs.microsoft.com/en-us/azure/databricks/administration-guide/access-control/tokens
    2. https://docs.microsoft.com/en-us/azure/databricks/dev-tools/api/latest/aad/service-prin-aad-token
9. load a Data Frame with sensitive information &
10. write encrypted data to tables or Parquet files &
11. manage sensitive information
    1. https://databricks.com/blog/2020/11/20/enforcing-column-level-encryption-and-avoiding-data-duplication-with-pii.html
    2. https://databricks.com/session_na20/encryption-and-masking-for-sensitive-apache-spark-analytics-addressing-ccpa-and-governance

4. **Monitor and Optimize Data Storage and Data Processing (10-15%)**
    1. Monitor data storage and data processing
        1. implement logging used by Azure Monitor
            1. https://docs.microsoft.com/en-us/azure/azure-monitor/logs/data-platform-logs
        2. configure monitoring services
            1. https://docs.microsoft.com/en-us/azure/azure-monitor/deploy
        3. measure performance of data movement
            1. https://docs.microsoft.com/en-us/azure/azure-sql/database/monitoring-with-dmvs
        4. monitor and update statistics about data across a system
        5. monitor data pipeline performance
            1. https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor
        6. measure query performance
            1. https://docs.microsoft.com/en-us/azure/azure-sql/database/query-performance-insight-use

   7. monitor cluster performance
      1. https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-key-scenarios-to-monitor
      2. https://docs.microsoft.com/en-us/azure/synapse-analytics/monitoring/how-to-monitor-using-azure-monitor
      3. https://docs.microsoft.com/en-us/azure/architecture/databricks-monitoring/
   8. understand custom logging options
      1. https://docs.microsoft.com/en-us/azure/azure-monitor/agents/data-sources-custom-logs
   9. schedule and monitor pipeline tests
   10. interpret Azure Monitor metrics and logs
       1. https://docs.microsoft.com/en-us/azure/azure-monitor/essentials/data-platform-metrics
   11. interpret a Spark directed acyclic graph (DAG)
2. Optimize and troubleshoot data storage and data processing
   1. compact small files
   2. rewrite user-defined functions (UDFs)
   3. handle skew in data
      1. https://en.wikipedia.org/wiki/Skewness
      2. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute#choose-a-distribution-column-with-data-that-distributes-evenly
      3. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute#determine-if-the-table-has-data-skew
   4. handle data spill
      1. https://en.wikipedia.org/wiki/Data_breach
      2. https://docs.microsoft.com/en-us/compliance/regulatory/gdpr-breach-notification
      3. https://docs.microsoft.com/en-us/compliance/regulatory/gdpr-breach-azure-dynamics
   5. tune shuffle partitions
      1. https://docs.microsoft.com/en-us/azure/architecture/databricks-monitoring/performance-troubleshooting
   6. find shuffling in a pipeline
   7. optimize resource management
   8. tune queries by using indexers
      1. https://docs.microsoft.com/en-us/azure/azure-sql/database/automatic-tuning-overview
      2. https://docs.microsoft.com/en-us/sql/relational-databases/automatic-tuning/automatic-tuning?view=sql-server-ver15
   9. tune queries by using cache

    1. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-caching

10. optimize pipelines for analytical or transactional purposes
11. optimize pipeline for descriptive versus analytical workloads
12. troubleshoot a failed spark job
    1. https://docs.microsoft.com/en-us/azure/databricks/kb/jobs/
    2. https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-known-issues
    3. https://docs.microsoft.com/en-us/azure/data-factory/data-factory-troubleshoot-guide
13. troubleshoot a failed pipeline run
    1. https://docs.microsoft.com/en-us/azure/data-factory/data-factory-troubleshoot-guide