BPP
UNIVERSITY
SCHOOL OF TECHNOLOGY

Level 5
Data Engineer

Higher Apprenticeship

Module 2

Databases and Data Lakes

Module Handbook
2024

MODULE OVERVIEW

Databases and Data Lakes

Module	Phase	Weekly Topics	Description
Data Fundamentals	Month 1	**1. Storing and Exploring Data** • **Overview and Importance**: ○ Understand why businesses retain data due to the benefits of cheap storage and Big Data utilisation. ○ Explore the strategic significance of data retention for businesses. • **Data Storage Techniques**: ○ **File Systems**: ▪ Storing files using HDFS (e.g., for storing CSV files), Key-Value stores, and Columnar formats such as Parquet. ▪ Traditional hierarchical storage using file systems and their limitations. ○ **Data Storage Options**: ▪ Definitions, examples, and comparisons of Data Lakes, Data Warehouses, and Operational Data Stores. ▪ Pros and cons of on-premise versus cloud storage options, with examples of cloud providers. • **Data Structuring Concepts**: ○ Introduction to data structuring concepts: Schema-on-Write and Relational Modelling. ○ Transition from traditional hierarchical storage using file systems to relational data models. ○ Essentials of relational databases: tables, schemas, and relationships. ○ Practical examples illustrating the transition from file-based storage to relational databases.	This module explains the importance of structuring of upstream data for downstream consumption, while making sure that data does not get lost or corrupted. Apprentices will learn about how data is stored and queried using filesystems, SQL and NoSQL, while being exposed to key database types, storage file formats and technical processes for data integration. We introduce key databases concepts like data normalisation and query optimisation, and the fundamental theory behind data consistency and availability to ensure that data is updated and accessed in guaranteed and reliable ways. You will learn skills of practically managing and monitoring the performance databases and data lake components, comprehend the significance of data quality, common database-related issues and their impact on data analysis. We focus on open-source standards throughout. Apprentices will be

		• **Data Lakehouse Architecture**: ○ Understanding the architecture and benefits of Data Lakehouses. • **Hands-On SQL Practice**: ○ Developing SQL skills through hands-on DDL (Data Definition Language) and DML (Data Manipulation Language) exercises. ○ Introduction to querying data with SQL on platforms such as AWS. **2. Schemas and Integration** □ **Relational Data Model**: • **Types of Keys and Relationships**: ○ Explore primary and foreign keys, and their role in establishing relationships in databases. ○ Discuss the history and evolution of the relational data model. • **Schemas and Metadata**: ○ Explain different types of schemas (e.g., star, snowflake) and their uses. ○ Discuss the importance of metadata in managing data, and how schemas are part of metadata. □ **Data Warehouses and Data Marts**: • **Distinctions and Use Cases**: ○ Understand data warehouses and data marts. ○ Distinguish between OLAP and OLTP systems and their relevance in business scenarios. □ **SQL Database Design and Integration**:	introduced to database performance monitoring. Data quality and performance EPA preparation receives emphasis in this module. You will also understand what regulatory considerations apply and how to keep your environment secure. This module also highlights the critical role of schemas and their integration into business systems. It covers advanced data modelling, including keys, relationships, and schema evolution, along with data warehouses, data marts, OLAP, and OLTP systems. The module also explores star schemas, data integration techniques, SQL database design, metastores, and tools like Apache Iceberg and AWS Glue. In a data-driven enterprise, understanding the complexities and capabilities of SQL is essential. This module explores the impacts of upstream data structuring on downstream applications, the ACID principles, and the importance of transactions in business use cases. Learners will develop skills in writing nested CRUD SQL queries, performing joins, carrying out

		• **Data Integration Techniques**: o Learn about star schemas, data integration techniques, and SQL database design. o Understand metastores and open table formats like Apache Iceberg. • **Comparison of SQL Databases**: o Highlight differences between SQL vendors (e.g., PostgreSQL, Amazon Redshift, MySQL, Microsoft SQL) and their impact on data architecture. o Discuss SQL standards and vendor-specific extensions. □ **Schema Evolution**: • **Managing Schema Changes**: o Discuss schema evolution and its necessity as business requirements change. o Explain how to manage schema changes without disrupting operations. o Provide examples of tools that support schema evolution, including Apache Iceberg. □ **Practical Exercises**: • **Designing a Star Schema**: o Guide through the practical design of a star schema using a hypothetical dataset. • **Using AWS Glue**: o Explore AWS-specific SQL features and integration with AWS Glue. • **Integrating New Data Sources**:	database normalisation, and applying groupBy, aggregation, and windowing techniques. Additionally, it covers appraising SQL database risks, disaster recovery strategies, and focuses on data quality and performance through portfolio work.

    - ○ Discuss challenges and strategies for maintaining data integrity during integration.
    - ○ Mention Snowflake for its cloud data warehousing capabilities.

□ **Vendor Differences and SQL Standards**:

- **Implications of Vendor-Specific SQL Extensions**:
    - ○ Explore variations in SQL standards across different databases.
    - ○ Discuss the impact of these differences on data migration and system integration.

□ **Open Source vs. Proprietary Software**:

- **Impact on Scalability, Cost, and Support**:
    - ○ Discuss how the choice of open source software affects scalability, cost, and community support.

3. Advanced SQL

- **Impact of Data Structuring**:
    - Explore how upstream data structuring influences downstream applications.
    - Understand the implications of data integrity and consistency.

□ **ACID Principles**:

- Detailed explanation of ACID (Atomicity, Consistency, Isolation, Durability).
- Importance of transactions in business use cases to ensure reliable operations.

		□ **SQL Query Skills**: • Writing nested CRUD (Create, Read, Update, Delete) SQL queries. • Performing various types of joins (inner, outer, left, right). • Implementing database normalisation to organise data efficiently. □ **Advanced SQL Techniques**: • Applying groupBy, aggregation, and windowing functions. • Techniques for optimising query performance and reducing execution time. □ **Risk Management and Disaster Recovery**: • Appraising SQL database risks and developing disaster recovery strategies. • Ensuring business continuity through robust recovery plans. □ **Portfolio Work**: • Focus on data quality and performance through practical exercises. • Emphasis on real-world applications and scenarios. **4. Database Administration and Optimisation** ● **Monitoring Database Operations**: • Techniques for effective database monitoring and ensuring operational efficiency. • Understanding key performance indicators (KPIs) and metrics. □ **Database and Query Performance**:	

- Explaining factors that affect database and query performance.
- Methods for diagnosing and resolving performance issues.

□ **Security and Regulatory Considerations**:

- Understanding the importance of database security.
- Compliance with regulatory standards and best practices for data protection.

□ **Operational Costs and Sustainability**:

- Explaining the costs associated with database operations.
- Strategies for achieving sustainability and cost-efficiency in database management.

□ **Optimisation Techniques**:

- Techniques for optimising SQL queries to improve performance.
- Understanding and implementing partitioning and sharding.

□ **Project Work**:

- Practical exercises focusing on timeliness and accountability in database administration.

5. NoSQL Fundamentals

- **Understanding NoSQL Databases**:

		• Introduction to NoSQL and its significance in handling unstructured and semi-structured data. • Overview of schema-on-read, BASE principles, and the CAP theorem. □ **Types of NoSQL Databases**: • Detailed exploration of different NoSQL data models: key-value, document, column-family, and graph databases. • Specific use cases and advantages of each NoSQL type. □ **Discovering MongoDB**: • Introduction to MongoDB and its core features. • Practical applications of MongoDB Atlas for cloud-based NoSQL solutions. □ **Data Transformation**: • Transitioning from unstructured data to MongoDB and SQL. • Techniques for converting and integrating data across different database systems. □ **Key Concepts and Technologies**: • Understanding BSON and Parquet formats. • Exploring NewSQL as a hybrid approach combining SQL and NoSQL benefits. □ **Project Deliverables**: • Overview of project deliverables, including practical exercises and presentations.	

		• Emphasis on applying NoSQL concepts in real-world scenarios.	

Outcomes

After finishing this module, you will be able to meet the following outcomes and KSBs:

- **Apply** standard industry tools and best practices for designing, maintaining and optimising the performance and security of databases, data lakes, warehouses and lakehouses, including identifying relevant open-source standards for storing data.
    - K3, K9, K11, K14, K20,
    - S3, S4, S6, S12, S14, S24, S25
    - B5, B6

- **Explain** the fundamental concepts of SQL and NoSQL approaches to storing, updating and querying data and the impacts of changes in upstream database systems on the needs of diverse data consumers, and **list** the processing and storage costs involved.
    - K2, K12, K17, K19, K24, K25, K26
    - S6, S9, S12, S23, S24
    - B2

- **Evidence** the use of schemas, metadata, data modelling, transactions and identifying, monitoring and optimising key data access metrics to guarantee the reliability and sustainability of data and mitigate data risks.
    - K1, K5, K7, K15, K24, K25
    - S3, S5, S6, S7, S10, S13, S27
    - B4

- **Appraise** user and business requirements to **organise** upstream data for accuracy, completeness, consistency, timeliness and accessibility to downstream consumers whilst **acting proactively** taking accountability to ensure deadlines are met.
    - K4, K9
    - S1, S2, S5, S13, S26, B1

Introduction to the Module

This comprehensive module underscores the importance of structuring upstream data for downstream consumption, ensuring that data integrity is maintained, and loss or corruption is prevented. Apprentices will delve into how data is stored and queried using filesystems, SQL, and NoSQL, gaining exposure to key database types, storage file formats, and technical processes for data integration. Key database concepts such as data normalisation and query optimisation are introduced, alongside fundamental theories of data consistency and availability, to ensure data is reliably updated and accessed.

Learners will acquire practical skills in managing and monitoring the performance of databases and data lake components, understanding the significance of data quality, addressing common database-related issues, and their impact on data analysis. Emphasis is placed on open-source standards, with apprentices being introduced to database performance monitoring. The module places significant focus on preparing for EPA (End-Point Assessment) related to data quality and performance, while also covering regulatory considerations and strategies for maintaining a secure environment.

Learners will gain essential skills in data normalisation, query optimisation, and database administration, preparing them for real-world applications and EPA. The module emphasises data quality, regulatory compliance, and security, equipping apprentices with the knowledge and skills to design and manage robust, reliable data architectures. By the end of this module, apprentices will be well-prepared to navigate the complexities of the data landscape, ensuring data integrity and driving value for businesses through effective data management practices.

Topic Descriptions:

Storing and Exploring Data

This topic explores why businesses retain data due to cheap storage and the advantages of Big Data. It covers various data storage techniques including traditional hierarchical file systems, HDFS, Key-Value stores, and columnar formats like Parquet. Learners will compare Data Lakes, Data Warehouses, and Operational Data Stores, and examine the pros and cons of on-premise versus cloud storage options. The journey from traditional hierarchical storage to relational data models is detailed, including Schema-on-Write and Relational Modelling concepts. Practical exercises in SQL on platforms like AWS will solidify these concepts.

Schemas and Integration

Schemas and their integration into business systems are crucial in data management. This topic delves into advanced data modelling, keys, relationships, schema evolution, and their impact on augmented analytics. Learners will gain a thorough understanding of data warehouses, data marts, and the distinctions between OLAP and OLTP systems. It covers star schemas, data integration techniques, SQL database design, metastores, and tools like Apache Iceberg and AWS Glue. Practical exercises in designing star schemas and integrating new data sources enhance learning.

Advanced SQL

Understanding SQL's complexities and capabilities is essential in a data-driven enterprise. This topic covers the impact of upstream data structuring on downstream applications, ACID principles, and the importance of transactions. Learners will develop skills in writing nested CRUD SQL queries, performing joins, normalising databases, and applying groupBy, aggregation, and windowing techniques. The topic also includes appraising SQL database risks, disaster recovery strategies, and focuses on data quality and performance through portfolio work.

Database Administration and Optimisation

This topic addresses monitoring database operations, understanding database and query performance, and addressing security and regulatory considerations. Learners will list relevant administration-related metrics, explain operational costs and sustainability, and understand processing and storage costs involved in parallelism. Techniques for optimising SQL queries, partitioning, and sharding are covered. Project work focuses on timeliness and accountability in database administration.

NoSQL Fundamentals

The fundamentals of NoSQL databases are introduced, including schema-on-read concepts, BASE principles, CAP theorem, and various NoSQL data models. Learners will explore MongoDB Atlas, NewSQL, and transitioning unstructured data to MongoDB and SQL. The topic covers different types of NoSQL databases, their specific use cases, and an overview of project deliverables. Practical exercises on using MongoDB and integrating new data sources provide hands-on experience.

Mode of Assessment

END POINT ASSESSMENT (EPA)

It is **important** to read all the assessment guide documents contained in the Programme Handbook, as they contain important details.

Reminder: Refer to the programme handbook for further guidance.

Core Reading:

Campbell, L., Majors, C. (2017). Database Reliability Engineering, O'Reilly Media [20 hours]

Done, P., Kamsky, A. (2023). Practical MongoDB Aggregations. Packt Publishing [25 hours]

King, T., Schwarzenbach, J., (2020). Managing Data Quality: A practical guide. British Computer Society [15 hours]

Perkins, L., Redmond, E., Wilson, J. (2018) Seven Databases in Seven Weeks. O'Reilly [25 hours]

Simsion, G., Witt, G. (2005) Data Modeling Essentials 3rd ed., Elsevier [20 hours]

Zhao, A. (2021). SQL Pocket Guide: A Guide to SQL Usage. O'Reilly Media [20 hours]