**AWS Cloud Analytics Migration Report**

**Project: Migrating On-Premises Power BI Solution to AWS Cloud  
Prepared by: Aman Singh  
Date: 15/09/2025**

**Executive Summary**

The pharmaceutical industry increasingly relies on data-driven decision-making to optimize operations, improve patient outcomes, and comply with regulatory requirements. Traditionally, the company relied on an on-premises Power BI solution to analyze sales, inventory, and forecasting data. While Power BI provided a robust analytics platform, it faced challenges with scalability, performance, and maintenance costs, especially with increasing data volumes and multiple business units.

The purpose of this project was to migrate the existing Power BI solution to AWS Cloud using cloud-native services. The goal was to maintain equivalent functionality while leveraging serverless architecture, automation, and cost-efficient cloud services. The solution architecture consists of:

* Amazon S3 as a scalable data lake for raw, processed, and presentation datasets.
* AWS Glue ETL jobs for data transformation, cleaning, and validation.
* Amazon Athena for SQL-based queries on S3 data.
* Amazon QuickSight for interactive dashboards and reporting.

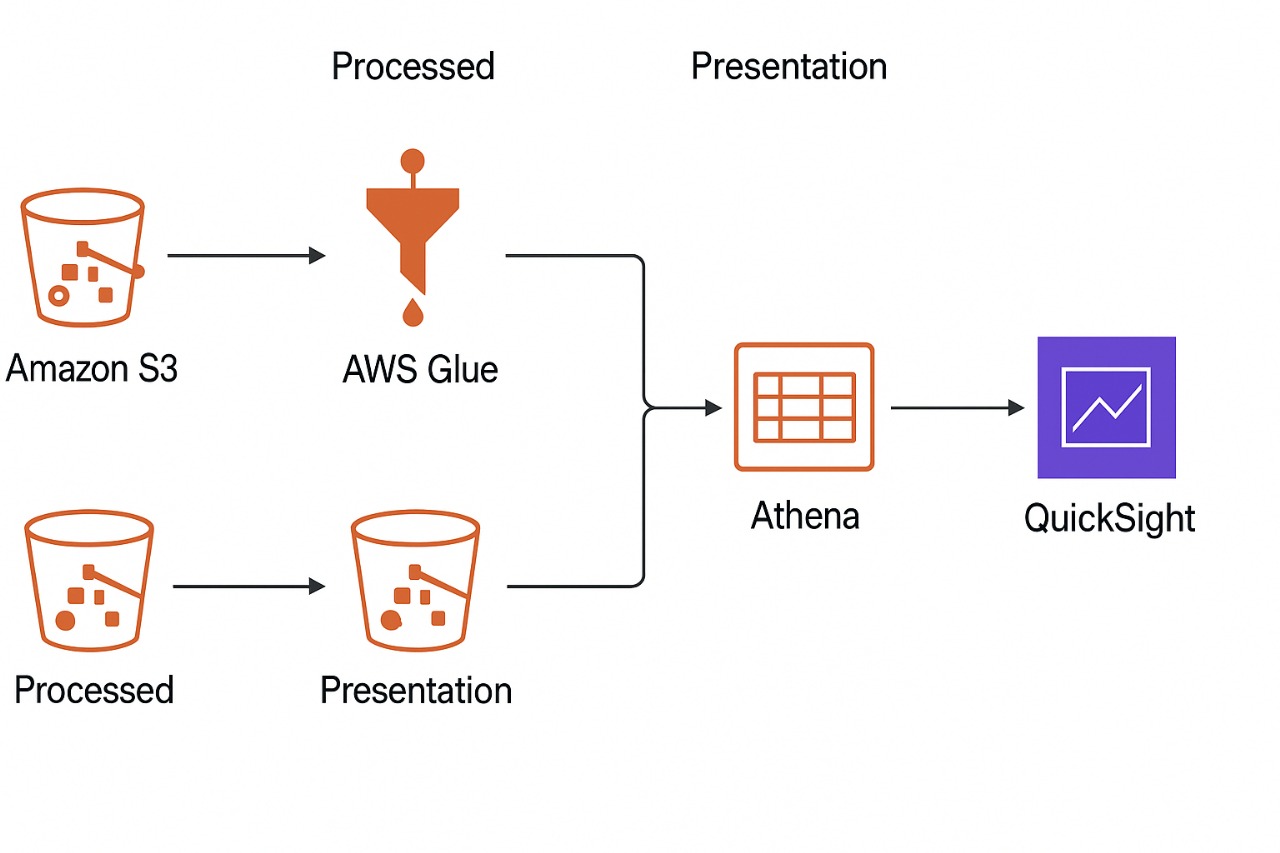
Key outcomes of this migration include:

1. Scalable and automated analytics pipeline that handles large datasets efficiently.
2. Presentation-ready datasets in Parquet format for faster querying and reduced cost.
3. QuickSight dashboards providing insights into sales, inventory, and product expiry, comparable to Power BI functionality.
4. Documented learning journey, cost analysis, and migration recommendations to guide future analytics initiatives.

The solution also incorporates best practices for security, governance, and compliance, including HIPAA considerations for sensitive pharmaceutical data.

**AWS architecture diagram:**

**Showing Raw → Processed → Presentation → QuickSight dashboards**



**1. Learning and Planning (Phase 1)**

Before starting the migration from on-premises Power BI to AWS Cloud, it was crucial to understand the AWS ecosystem, cloud-native services, and their integration capabilities. In this phase, the focus was on comprehending the architecture of a data lake, cloud analytics pipelines, and business intelligence tools offered by AWS.

AWS QuickSight was analyzed as the primary BI tool for this migration. QuickSight offers a cloud-native platform with SPICE (Super-fast, Parallel, In-memory Calculation Engine), enabling high-speed analytics on large datasets. Unlike Power BI, which primarily relies on either local data sources or dedicated servers, QuickSight can connect directly to Amazon S3, Athena, Redshift, and other AWS data sources without complex connectivity setups. During the learning phase, the functionality of QuickSight calculated fields, drill-downs, filters, and visualizations was explored. For example, calculated fields allow creating metrics such as inventory turnover or profit margins, similar to DAX formulas in Power BI. Drill-downs enable users to analyze data hierarchically, such as region → product category → product, providing interactive insights.

Understanding Amazon S3 as a data lake was equally important. S3 provides virtually unlimited storage, high durability, and cost efficiency, making it suitable for storing both raw and processed pharmaceutical data. Best practices were studied, including organizing the data lake into raw, processed, and presentation layers. The raw layer stores unprocessed source files for traceability and auditing. The processed layer contains cleaned and transformed datasets in a columnar format such as Parquet, which improves query performance in Athena and reduces storage costs. The presentation layer consists of aggregated datasets optimized for dashboards and business analytics. Partitions, such as year/month/day for sales data or region/product\_category for inventory, were considered to reduce query costs and improve efficiency.

AWS Glue was explored as the serverless ETL service. Glue allows for automation in data cleaning, transformation, and cataloging. Crawlers detect schema changes automatically, while jobs can handle joining multiple tables, data type conversions, null handling, and duplicate removal. The importance of job logging for monitoring and debugging was highlighted during this phase.

Finally, a comparative analysis of QuickSight and Power BI was conducted. While Power BI has advanced visualizations and DAX functionality, QuickSight offers native AWS integration, cost efficiency, and serverless scalability. Challenges such as reimplementing complex DAX formulas and optimizing SPICE datasets were anticipated.

Planning also involved defining migration requirements, such as datasets needed, transformation logic, KPI calculations, and compliance considerations (HIPAA). A detailed roadmap was drafted to guide subsequent ETL, dashboard development, and governance phases.

**2. Data Lake & ETL Implementation (Phase 2)**

After completing the learning and planning phase, the next step was to implement the data lake architecture and ETL pipelines on AWS. The goal was to transform raw pharmaceutical data into structured, analytics-ready datasets while ensuring scalability, cost efficiency, and automation.

Amazon S3 Data Lake Design formed the foundation of this implementation. The data lake was structured into three main layers:

1. *Raw Layer*: All source files, including sales transactions, product details, and inventory data, were ingested into the raw S3 bucket. This layer preserves the original data for traceability, auditing, and potential reprocessing in the future. Files were organized by logical categories such as s3://as-pharmadl-raw/sales/, s3://as-pharmadl-raw/inventory/, and s3://as-pharmadl-raw/products/.
2. *Processed Layer*: Data cleaning and transformations were performed in this layer using AWS Glue ETL jobs. The processed data was stored in Parquet format to leverage columnar storage benefits for cost reduction and query performance. Partitioning was implemented for large datasets; for instance, sales data was partitioned by year/month/day, while inventory was partitioned by region/product\_category. This significantly reduced Athena query costs and improved execution speed.
3. *Presentation Layer*: The processed data was further aggregated to create presentation-ready datasets. Tables such as sales\_summary, inventory\_turnover, and expiry\_alerts were generated for direct consumption in QuickSight dashboards. This layer prioritized performance and simplicity for end-user analytics.

AWS Glue ETL Jobs were critical for the data transformation pipeline. Several jobs were created to handle different datasets. The ETL process included:

* *Data Cleaning*: Removing duplicates, handling null values, converting data types for consistency, and standardizing product and region names.
* *Transformations*: Joining multiple datasets, such as linking sales transactions with product details and inventory levels, to create consolidated analytics-ready tables.
* *Aggregations*: Calculating key metrics like total sales per product, average stock levels, inventory turnover ratios, and identifying products nearing expiry within six months.

Glue Crawlers were configured to automatically detect schema changes and populate the Glue Data Catalog, which served as the metadata repository for Athena queries. Logging was enabled in each job to capture execution status, errors, and performance metrics. This provided visibility into pipeline health and simplified debugging.

Athena Queries were used to create the presentation datasets. For example:

* *Sales Summary*: Aggregated total sales per product.
* *Inventory Turnover*: Calculated as the ratio of total sales to average stock.
* *Expiry Alerts:* Selected products with expiry dates within six months.

Using Athena in combination with Glue ensured serverless scalability, meaning that the system could process millions of records without managing servers or provisioning infrastructure. Partitioned Parquet datasets allowed for efficient queries and minimized costs because Athena charges per amount of data scanned.

Automation was implemented using CloudWatch Events, which triggered Glue jobs on a daily schedule. This ensured that the processed and presentation datasets remained up-to-date without manual intervention. By integrating S3, Glue, and Athena, the solution created a fully automated, cloud-native ETL pipeline that mirrored the functionality of traditional on-premises Power BI workflows but with higher efficiency, flexibility, and scalability.

*Challenges and Solutions:*

* *Schema Changes*: Some datasets had evolving schemas, causing Glue jobs to fail. Using dynamic frames in Glue allowed the ETL jobs to adapt to schema changes.
* *Large Dataset Performance*: Initial queries scanned large amounts of data. Partitioning and using Parquet files significantly reduced scan sizes.
* *Error Tracking*: Glue job errors were difficult to trace without logging. Enabling CloudWatch logs provided detailed error messages, improving debugging efficiency.

In conclusion, this phase successfully established a robust, automated data lake and ETL pipeline, transforming raw pharmaceutical datasets into clean, structured, and analytics-ready tables. This foundation is essential for building reliable dashboards and ensuring data-driven decision-making in the cloud environment.

**3. QuickSight Dashboards & Analytics (Phase 2 Cont.)**

Once the data lake and ETL pipelines were fully implemented, the next phase focused on visualizing the processed data using AWS QuickSight. The goal was to replicate the functionality of the on-premises Power BI dashboards while leveraging QuickSight’s cloud-native capabilities such as SPICE, interactive filters, drill-downs, and calculated fields.

**3.1 Dataset Integration in QuickSight**

The first step was connecting QuickSight to the presentation layer datasets stored in S3 and cataloged in Athena. Three key datasets were imported:

1. *Sales Summary* – Containing aggregated total sales per product.
2. *Inventory Turnover* – Calculated as total sales divided by average stock, providing insight into product movement.
3. *Expiry Alerts* – Listing products with expiry dates within six months, critical for regulatory compliance and inventory planning.

QuickSight datasets were created in SPICE to improve performance and reduce latency for interactive dashboards. SPICE allows in-memory caching of datasets, enabling users to perform rapid queries and filter operations even on large tables.

**3.2 Dashboard Design and Visualization**

**Sales Performance Dashboard:**

* *Objective*: Monitor product-level revenue and regional trends.
* *Visualizations*: Bar charts for top-selling products, line charts for month-over-month sales, and geographic heatmaps for regional performance.
* *Features*: Drill-downs enabled analysis from regional summaries down to individual product sales. Filters allowed users to segment data by product category or time period.

**Inventory Analysis Dashboard:**

* *Objective*: Provide insight into inventory efficiency and potential stockouts.
* *Visualizations*: Tables for current stock levels, line charts for inventory turnover, and conditional formatting to highlight low-stock products.
* *Calculated Fields*: Metrics such as inventory turnover (sales ÷ average stock) were implemented directly in QuickSight, replicating the logic previously calculated in Athena.

**Expiry Alerts Dashboard:**

* Objective: Highlight products nearing expiry for proactive management.
* *Visualizations*: Color-coded tables and heatmaps indicated products at risk of expiry within the next six months.
* *Automation*: Dashboards were connected to SPICE datasets updated daily through the Glue ETL pipeline, ensuring real-time reflection of changes in source data.

**3.3 Advanced Features and Interactivity**

QuickSight’s interactive features allowed user-driven analysis:

* *Filters and Parameters*: Users could filter by region, product category, or time frame, providing customizable views of data.
* *Drill-Downs*: Enabled multi-level exploration, such as region → warehouse → product → day-level sales.
* *Conditional Formatting*: Highlighted KPIs that required immediate attention, e.g., low inventory or products nearing expiry.
* *Calculated Fields*: Allowed on-the-fly metric computation, such as profit margins, inventory turnover rates, or percentage of total sales.

**3.4 Automation and Refresh**

Dashboards relied on daily ETL jobs to refresh data. QuickSight SPICE datasets were scheduled to update automatically after Glue ETL completion, ensuring that users always accessed the latest data without manual intervention. This workflow replicated the scheduled refresh functionality of Power BI in a serverless cloud environment**.**

**3.5 Security and Access Control**

* *IAM Integration*: Access to datasets and dashboards was controlled via IAM roles, ensuring only authorized personnel could view or edit data.
* *Row-Level Security (RLS):* Implemented to restrict sensitive pharmaceutical information to relevant users, in compliance with HIPAA regulations.
* *Audit Trails*: QuickSight activity was monitored using CloudTrail to log user interactions, data access, and dashboard modifications.

**3.6 Challenges and Solutions**

* *Translating Power BI Dashboards*: Some advanced DAX formulas required re-implementation using QuickSight calculated fields and Athena SQL queries.
* *Performance with Large Datasets*: Direct queries on large tables initially caused delays. The use of SPICE datasets and partitioned Parquet files resolved this issue.
* *Dynamic Reporting Requirements*: Frequent changes in business logic, such as adding new KPIs or adjusting filters, were addressed by modular dashboard design and reusable calculated fields.

**3.7 Summary**

This phase successfully demonstrated that AWS QuickSight can replicate and enhance Power BI functionality in a cloud-native environment. Dashboards are interactive, automated, and scalable, providing real-time insights into sales, inventory, and regulatory compliance. By leveraging SPICE, calculated fields, drill-downs, and conditional formatting, the dashboards offer a rich, user-friendly experience while ensuring compliance, security, and efficiency.

**4. Analysis, Cost, Governance & Recommendations (Phase 3)**

The final phase of the project focused on analyzing the implemented AWS analytics solution, evaluating costs, ensuring governance and compliance, and providing recommendations for future improvements. This phase consolidates lessons learned from Phases 1 and 2, highlights scalability and performance considerations, and compares AWS cloud analytics with the existing on-premises Power BI solution.

**4.1 Performance and Analysis**

The implemented AWS solution was tested with sample pharmaceutical datasets to evaluate its performance. The ETL pipelines using AWS Glue processed millions of records efficiently due to serverless scaling. Partitioned Parquet files significantly reduced query time in Athena. For example, querying sales data for a specific month scanned only relevant partitions, reducing both execution time and cost.

QuickSight dashboards were evaluated for interactivity and user experience. Drill-downs allowed users to explore high-level summaries down to individual product transactions. Filters enabled slicing and dicing by region, product category, or time period, providing insights similar to Power BI dashboards. Calculated fields allowed on-the-fly metrics like inventory turnover, profit margins, and products nearing expiry, enhancing decision-making capabilities.

The analytics solution also facilitated proactive management. For instance, the expiry alerts dashboard enabled warehouse managers to identify products approaching expiration, reducing waste and regulatory risk. Inventory turnover analysis helped supply chain managers optimize stock levels, ensuring better availability while minimizing overstocking costs.

**4.2 Cost Analysis**

A key benefit of migrating to AWS was cost optimization.

* *Amazon S3:* Data stored in Parquet format with partitions reduced storage costs compared to uncompressed CSV files.
* *AWS Glue*: Serverless ETL jobs billed per DPU-hour allowed scaling only when needed, eliminating the need for permanent on-premises ETL infrastructure.
* *Amazon Athena*: Pay-per-query model ensured costs were incurred only when data was analyzed. Partitioning further minimized scanned data, reducing costs.
* *QuickSight*: SPICE datasets reduced compute usage while enabling fast interactive dashboards. The pay-per-session model is cost-effective for organizations with multiple users who access dashboards intermittently.

Comparatively, an on-premises Power BI solution would require significant investment in server infrastructure, maintenance, licensing, and manual data refreshes, making AWS a more scalable and cost-effective solution in the long term.

**4.3 Security and Governance**

Governance and compliance were critical due to the sensitive nature of pharmaceutical data. Several measures were implemented:

* *IAM Roles and Policies*: Controlled access to S3 buckets, Athena queries, Glue jobs, and QuickSight dashboards. Only authorized users could perform specific actions.
* *Encryption*: All data was encrypted at rest in S3 and in transit via TLS, ensuring data security.
* *Row-Level Security (RLS):* QuickSight dashboards implemented RLS to restrict data visibility to users based on roles, adhering to HIPAA compliance requirements.
* *Audit Trails and Monitoring*: AWS CloudTrail and CloudWatch were used to monitor data access, track ETL job execution, and capture logs for auditing purposes.

This approach ensures data integrity, confidentiality, and regulatory compliance while maintaining accessibility for authorized personnel.

**4.4 Migration Recommendations**

Based on the analysis and implementation experience, the following recommendations were made:

1. *Maintain a Structured Data Lake*: Keep raw, processed, and presentation layers separate to ensure data traceability and facilitate future analytics projects.
2. *Automate ETL and Refresh Pipelines*: Continue using Glue and Athena to schedule automatic data refreshes, ensuring dashboards always reflect the latest data.
3. *Implement Security Best Practices*: Continue enforcing encryption, IAM policies, RLS, and monitoring for governance and compliance.
4. *Leverage Advanced AWS Services*: Future integration with AWS Lambda can automate ETL on new data ingestion. SageMaker can be used for predictive analytics, such as forecasting demand or detecting anomalies in sales or inventory. Real-time analytics with Kinesis or MSK can enable immediate decision-making.

**4.5 Comparison with Power BI On-Premises**

The AWS solution provides several advantages over the traditional on-premises Power BI setup:

* *Scalability:* Can handle millions of records without manual server upgrades.
* *Cost Efficiency:* Pay-per-use billing reduces operational costs.
* *Automation*: ETL pipelines and SPICE datasets allow automatic updates, eliminating manual data refreshes.
* *Integration*: Native integration with AWS analytics and ML services enables advanced analytics and real-time insights**.**

While Power BI excels in advanced visualization capabilities, AWS QuickSight coupled with Glue, Athena, and S3 provides a robust, cloud-native solution that is scalable, automated, and cost-effective.

**4.6 Future Learning Roadmap**

To further enhance the analytics solution, future steps include:

* Implement Lambda triggers to automatically initiate ETL jobs on new S3 uploads.
* Use SageMaker to deploy machine learning models for predictive analysis on sales trends or inventory requirements.
* Explore real-time streaming analytics using AWS Kinesis to monitor inventory levels and sales in near real-time.
* Enhance QuickSight dashboards with embedded analytics for internal and external stakeholders.

**4.7 Summary**

This phase demonstrated that AWS provides a secure, scalable, and cost-effective analytics ecosystem for pharmaceutical data. By combining data lake architecture, ETL automation, serverless querying, and interactive dashboards, the solution not only replicates Power BI functionality but also offers enhanced scalability, automation, and integration possibilities. With proper governance and continuous improvement, this solution can serve as the foundation for advanced cloud analytics and machine learning initiatives.