Module Code: CT047-3-M-BDAT

Module Name: Big Data Analytics & Technologies

Individual Assignment - 2

A Conceptual Big Data Ecosystem for Environmental Analytics in Malaysia

Student Name: Muhammad Yousouf Ali Budullah
TP Number: TP086704
Intake Code: APDMF2501DSBA(BI)(PR)
Programme: MSc Data Science & Business Analytics
Module Lecturer: Assoc. Prof. Dr. V. Sivakumar

Date of Submission: June 2, 2025

Contents

1 Introduction

Malaysia's Department of Environment has encountered growing challenges in managing large, complex environmental data throughout fields like climate change, pollution and biodiversity. With current data collection methods dependent on periodic reporting and basic statistical tools, insights are frequently delayed or missing. The exigency of the situation requires real-time responses with a more advanced, integrated solution. This proposal describes a big data analytics framework to respond to the inefficiencies faced by taking advantage of available scalable technologies spanning ingestion, storage, processing, and analysis layers to promote faster decision making, improve accuracy and facilitate an intelligent environmental governance.

2 Benefits of Advanced Environmental Data Analysis

Advanced environmental data analysis provides progressive benefits for both public decision making and market flexibility amid growing climate uncertainty. By utilising big data technologies, the Department of Environment can transition from a reactive to proactive approach to environmental management. This assists in improved forecasting, smarter resource allocation and timely interventions, which are crucial in a country susceptible to floods, haze and biodiversity risks.

2.1 Applicable Analytics

Firstly, descriptive analytics plays an essential role by summarising historical environmental data. For instance, recurring patterns in daily air quality or annual haze indexes can be revealed and reported readily. This would allow stakeholders to briefly understand what has occurred over time. However, while useful for creating synopses, descriptive analysis is static and not actionable as it does not describe underlying causes.

Building upon this, diagnostic analysis aids in discovering why environmental incidents occur. It allows the Department of Environment to link outcomes with their cause, for example connecting river pollution with specific industrial zones. This supports root cause analysis and targeted interventions. Nonetheless, practical diagnostics require detailed and integrated datasets, which would be challenging due to the current systems fragmented nature.

Moving forward, predictive analytics emphasises forecasting future environmental scenarios using historical data patterns and real-time data. A few applications could include flooding forecasting based on rainfall trends or predicting air quality drops. This increases readiness and enables early warnings for civilians. Nevertheless, its accuracy is dependent on the model's quality and input data.

In continuation, prescriptive analytics presents actionable suggestions based on the patterns

discovered. This could include recommending stricter emission controls during critical periods or advising policymakers on useful environmental regulations. Despite its benefits, it also requires comprehensive modelling infrastructure and advanced analytical capabilities.

2.2 Impact on Decision-Making

In relation to public decision making, advanced analytics allows the Department of Environment to make faster, more accurate and context aware decisions. For example, real time data streaming from sensors can trigger an alert for unusual pollutant levels, enabling an immediate and proactive response to prevent environmental hazards. Policy development would also become more comprehensive, as trends and root cause analysis would provide a stronger basis for environmental regulations. Furthermore, prescriptive suggestions gathered from predictive models could assist personnel in deciding where to allocate limited resources, thus improving operational efficiency and public safety. The integration of advanced analytics would create a proactive rather than reactive governance, decreasing response time and improving accountability by the Department.

2.3 Impact on Market Stability

From a market perspective, environmental analysis plays a stabilising role throughout various sectors. In agriculture, predictive models could help farmers make better decisions about their crop cycles based on weather forecasts or soil data, reducing yield loss. For Insurance companies, they could assess environmental risk with greater accuracy, enabling them to create fair policies. The energy sector would profit from demand forecasting aligned with climate trends, improving load balancing on the electric grid and integration of renewable energy solutions. Tourism and urban development would also see gains from air and water quality monitoring which could shape infrastructure plans. By reducing uncertainty and improving prediction, analytics could provide flexible and strategic planning to both public and private sectors.

3 Big Data Ecosystem: Current Limitations and Proposed Architecture

3.1 Current Ecosystem Analysis and Limitations

Presently, the Department of Environment in Malaysia relies on traditional data collection and analysis methods. Based on the case study, data is gathered periodically from multiple disconnected sources and analysed using basic statistical tools. It is assumed that tools such as Excel, SQL databases, or Statistical Package for the Social Sciences (SPSS) are utilised for storing and analysis of environmental data. These methods are adequate for small-scale historical analysis;

however, they impose major limitations in a climate sensitive domain that requires real time insight and large scale integration.

Firstly, the dependence on periodic manual reporting introduces a delay between data collection and insight creation. For instance, pollution data from sensors or field reports could potentially take long time to be compiled, verified and analysed delaying immediate action. Secondly, the fragmentation of data throughout various sources like weather reports, pollution indices, and biodiversity records prevents the Department from performing cross-domain analysis or detect environmental trends. Furthermore, the traditional tools mentioned are all unable to process unstructured or semi-structured data types including satellite footage, sensor feeds or public complaints from social media.

Additionally, with the current infrastructure it is unrealistic for it to support automation, predictive modeling or even scalable storage. Without a consolidated data pipeline or cloud-based infrastructure, scaling operations to handle growing data volumes from climate sensors and Internet of Things (IoT) devices becomes more complicated. These flaws prevent not only the Department's efficiency but also limits its ability to provide timely insights to policymakers and the public. Given the limitations of the current ecosystem, the adoption of a big data so-

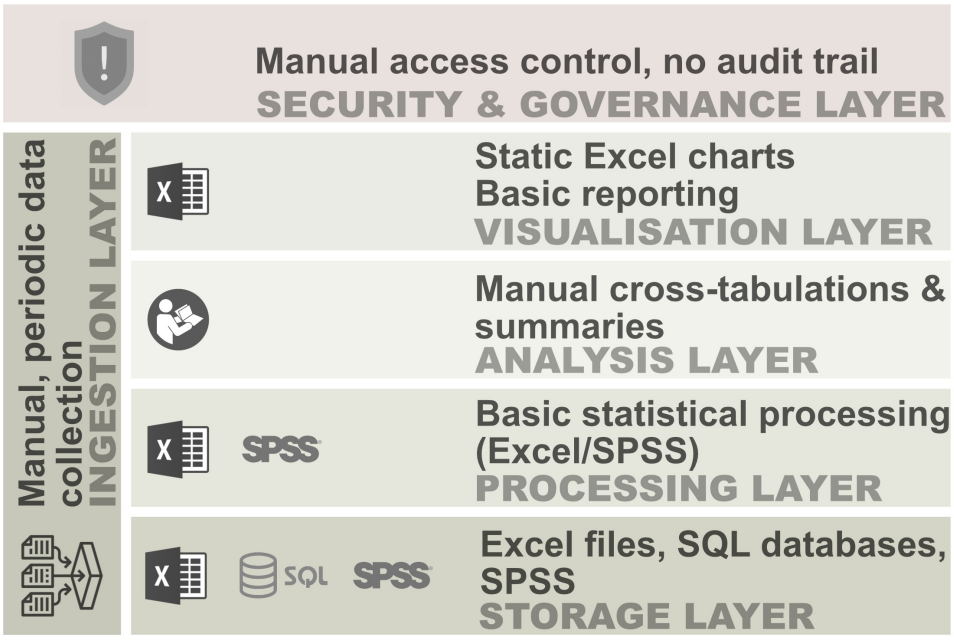

Figure 1: Current Ecosystem

lution is necessary. A comprehensive system would allow real time ingestion of multiple data sources, scalable storage for all data types (structured, unstructured and semi-Structured) and provide advanced processing capabilities for predictive and prescriptive analytics. This would not only improve response time but also support policy and improve the Department's long term environmental strategy. As climate events and become more frequent and unpredictable

it's necessary for effective environmental governance, and this can be achieve by a big data driven framework.

3.2 Conceptual Big Data Architecture

To overcome the limitations identified with the current system used by the Department of Environment's, a conceptual big data architecture is required. This architecture has been designed to be modular, scalable and adaptable to the various types of environmental data present. It facilitates efficient ingestion, storage, processing and analysis of both real time and historical data to support timely, data driven decision making.

The architecture follows a layered approach, where each layer manages a function of the big data pipeline. These layer cooperate together to generate actionable insight from the environmental data.

3.2.1 Data Ingestion Layer

The data ingestion layer is responsible for collecting environmental data from various sources. This includes IoT sensors, air and water quality monitors, satellite footage and the publics opinion from web forms or social media to name a few. It's imperative that the ingestion layer can handle both real time and batch ingestion to capture all data types, structure, unstructured and semi-structured.

3.2.2 Data Storage Layer

After the data has been ingested, it needs to be stored in a distributed and scalable system capable of managing multiple data types. The storage should enable long-term storing of raw environmental data whilst also facilitating fast access for processed dataset. Local on site and cloud storage solutions are viable and can be considered to balance performance, cost and compliance needs based on the department.

3.2.3 Data Processing Layer

The data processing layer serves as a transformative operation to create usable formats from the raw environmental data. It includes data cleaning, filtering and enhancement. The processing should support both batch processing such as a month's worth of CO2 emissions or stream processing like detecting a spice in CO2 emissions in real time. This makes sure that the Department can perform both historical and real time analytics.

3.2.4 Data Analysis Layer

Once the data has been processed, multiple analytical methods can be applied to derive insights from the data. These include descriptive analysis to understand current environmental data,

diagnostic analysis to find causes of anomalies in the data, predictive models for prediciting climate trends, and prescriptive analytics to recommend policies.

3.2.5 Data Visualisation Layer

The insights created from data anlyisis are delivered through dashboards, reports, and alert systems modified for different user needs. Visualisations should be insightful, real-time where necessary, and easily accessible to both technical and non-technical decision-makers within the Department of Environment.

3.2.6 Security and Governance Layer

Security and data governance span all layers. This includes data encryption, user access control, audit logging, and compliance to regulatory frameworks like Malaysia's PDPA. Proper governance guarantees that sensitive environmental and location-based data is protected while remaining accessible for appropriate analysis.

4 Adoption of Big Data Tools and Justification

4.1 Data Ingestion Layer

This layer takes data from multiple sources, including IoT sensors, satellite feeds, weather stations, and even social media. The tools featured below were chosen based on their ability to handle batch and real-time data in all forms.

4.1.1 Apache NiFi

Apache NiFi is an open source tool that automates the flow of data between various systems. It has a simple interface that allows users to create and maintain data pipelines without writing code. NiFi may filter, route, and transform data in real-time or batch mode, depending on how the system is configured ("Apache NiFi User Guide", n.d.). In Poland, the Chief Inspectorate of Environmental Protection employed NiFi to handle air quality data from sensors located around the country (Racka & Płocku, 2022). This makes it a good match for the Department of Environment, which likewise has to collect continuous environmental data from different sources and ensure that it is processed efficiently and safely.

4.1.2 Apache Kafka

Apache Kafka is a tool designed to process large amounts of data fast and reliably. It employs a publish and subscribe model, which enables several systems to submit and receive data independently. This makes Kafka a good option for large-scale implementations in which data

must transfer across systems without latency (“Apache Kafka”, n.d.). The UK Met Office utilizes Kafka to handle and transfer live data from weather stations to forecasting and analysis systems (Kenyon, 2017).This demonstrates that Kafka can help the Department achieve its goal of gathering real-time data from environmental sensors and guaranteeing that the data reaches the systems that require it without failure or delay.

4.1.3 Apache Flume

Apache Flume is a service that collects and transports data from several sources to a centralized location, typically for storage or future analysis. It is mostly used for event logs or semi-structured data obtained from sensors or field devices (“Flume 1.11.0 User Guide — Apache Flume”, n.d.). Flume was utilized in Barcelona’s smart city program to collect environmental data from air quality sensors and transfer it to storage for processing and monitoring (BU, Coforge-Salesforce, 2017). This type of arrangement is comparable to what the Department may want in rural or isolated places where data from sensors must be transferred reliably without the use of complex setups.

4.2 Data Storage Layer

After ingestion, the data needs to be stored securely and in a scalable way. These tools support different data types and allow both short-term access and long-term durability.

4.2.1 Apache HDFS

HDFS is a distributed file system that stores large amounts of data across multiple hardware. It is ideally suited for batch processing and offers fault tolerance via data replication (Apache Hadoop, n.d.). Its integration with tools such as Apache Spark and Hive makes it suitable for working with structured or semi-structured data. For the Department of Environment, HDFS would be ideal for storing long-term historical data such as pollution logs, emission records, and monthly reports. It works effectively in on-premise environments where the Department might prefer more control over its data and infrastructure rather than depending on third-party cloud providers.

4.2.2 Amazon S3

Amazon S3 is a cloud-based object storage solution designed for high durability, flexibility, and scalability. It can manage virtually any amount of data and supports a variety of formats such as JSON, CSV, images, and even binary sensor outputs (AWS, 2022). One of the primary reasons S3 is relevant here is its integration with cloud-native processing tools such as AWS Glue and SageMaker, which could facilitate any future analytics or machine learning initiatives that the Department wishes to pursue. Furthermore, S3’s lifespan regulations make it simple

to distinguish between live datasets and archived information, allowing for more effective cost management.

4.2.3 Google Cloud Storage

Google Cloud Storage is another cloud-based solution that offers scalable and highly available object storage. It provides real-time data access and works well with processing tools such as BigQuery and Dataflow ("Cloud Storage documentation", n.d.). One valuable feature is automatic class transfers based on usage, which can help minimize storage costs over time without requiring manual involvement. Google Cloud Storage is particularly valuable for geospatial data and climate-related imagery that the Department may collect for analysis or reporting purposes. With built-in encryption, role-based access, and audit logs, it also supports the security and compliance requirements of the public sector.

4.3 Data Processing Layer

This layer transforms raw data into clean, usable formats. The tools here are picked for their ability to handle both real-time streams and historical batch data, with strong automation support.

4.3.1 Apache Spark

Apache Spark is an open-source data processing engine designed to handle large batch and streaming workloads. It supports machine learning, SQL, and graph processing out of the box and is praised for its speed thanks to in-memory computation (Apache Spark, 2019). In a study on PM2.5 air pollution forecasting, researchers used Spark to create a predictive framework capable of processing huge sensor datasets and effectively running ensemble machine learning models (Shih et al., 2021). This demonstrates that Spark is suitable for the Department's long-term analytics activities, such as projecting emissions or analysing annual pollution trends, particularly when dealing with historical or aggregated data.

4.3.2 Apache Flink

Apache Flink is a real-time stream processing engine that can handle complicated events and create time-based windows("Apache Flink Documentation", 2024). It is designed to manage continuous data streams from systems such as environmental sensors and monitoring devices. In a recent study, Flink was employed in a framework that processed live IoT sensor data for smart environmental monitoring, delivering real-time alerts and analysis (Dinakar, 2024). This makes Flink appropriate for the Department's real-time monitoring requirements, such as warning when pollution standards are exceeded or rainfall levels rise unexpectedly.

4.3.3 Google Cloud Dataflow

Google Cloud Dataflow is a managed service that can handle both batch and streaming data processing. Its based the Apache Beam programming model and integrates nicely with other Google services like BigQuery and Cloud Storage ("Dataflow documentation", n.d.). Aclima, an air quality monitoring startup, uses Dataflow to handle and analyse billions of environmental data points collected from mobile and fixed-location sensors ("Aclima Case Study", n.d.). For the Department, Dataflow can assist manage large-scale sensor data in a flexible, cloud-based environment that scales automatically and eliminates the need to manage physical infrastructure.

4.4 Data Analysis Layer

This is where insights are generated. The tools listed can handle large structured datasets efficiently using SQL-like queries and have been used in similar environmental use cases.

4.4.1 Apache Hive

Apache Hive is a Hadoop-based data warehousing system that allows users to perform SQL-like searches on massive datasets stored in distributed storage systems ("Home - Apache Hive - Apache Software Foundation", n.d.). It has been used successfully in environmental data analysis. For instance, one study used Hive to analyse climate change data, collecting information from multiple sources and employing Hadoop's distributed processing capabilities to store and process the data in parallel. Hive was subsequently used to manage and analyse the data, which included performing complicated queries and developing visualisations (Greca et al., 2023). Hive can help the Department produce monthly or annual summaries of air quality and emissions data, which will aid in long-term environmental monitoring and reporting.

4.4.2 Google BigQuery

Google BigQuery is a fully managed, serverless data warehouse that allows for quick SQL searches by utilising Google Cloud's infrastructure ("BigQuery documentation", n.d.). It has been used in environmental initiatives such as the creation of a publicly accessible BigQuery collection of US National Water Model operating forecasts, which enables rapid data analysis and application development (Markert et al., 2024). BigQuery provides the Department with a scalable solution for analyzing large amounts of environmental data without the need to manage physical servers, allowing for more effective data processing and reporting.

4.4.3 Trino

Trino is a distributed SQL query engine that can query data from multiple sources without requiring data migration ("Trino documentation — Trino 475 Documentation", 2025). It has

been explored in environmental monitoring programs, such as creating streaming data pipelines to ingest and analyse data from the UK Environment Agency's sensor network, which reports on parameters such as river levels, rainfall, and temperature It has been used in environmental monitoring programs, such as creating streaming data pipelines to ingest and analyse data from the UK Environment Agency's sensor network, which reports on parameters such as river levels, rainfall, and temperature ("Data Exploration with Tableflow, Apache Iceberg, and Trino", 2025). Trino gives the Department the ability to query data stored in a variety of formats and locations, allowing for a comprehensive analysis of environmental data received from multiple sources.

4.5 Data Visualisation Layer

Visualisation helps make sense of the insights. These tools offer interactive dashboards and reports, suitable for both internal decision-making and public engagement.

4.5.1 Microsoft Power BI

Microsoft's Power BI is a business intelligence platform which allows users to create interactive dashboards and visual reports via a drag-and-drop interface. It can import data from multiple sources, including Excel, SQL databases, APIs, and cloud services("Microsoft Industry Clouds", n.d.). Various environmental and government agencies utilise Power BI to track indicators such as water quality, garbage levels, and air pollution. An instance is the UK Department for Environment, Food, and Rural Affairs (DEFRA), which uses Power BI to visualize environmental data such as pollution exposure across regions (DEFRA, 2019). For the Department, Power BI offers a familiar, user-friendly interface for both technical and non-technical people to access and comprehend environmental data patterns.

4.5.2 Tableau

Tableau is another popular visual analytics tool that allows users to analyse data via interactive charts, maps, and dashboards (Tableau, 2023). It is well-known for its powerful geospatial capabilities, which are especially helpful when plotting environmental data such as pollution levels, rainfall distribution, or emission zones on a map. One example of its use, the State of New York used Tableau Public to give real-time dashboards on environmental sustainability initiatives such waste management ("Composting - Solid Waste Management Facilities Map", n.d.). Tableau would enable the Department to create visual reports for internal use or public engagement, which encourages transparency and evidence-based decision making.

4.6 Security and Governance Layer

This layer ensures the right people have access to the right data, while staying compliant with laws like PDPA. The tools chosen here are built for managing access control and audit logging.

4.6.1 Apache Ranger

Apache Ranger is used in Hadoop systems to control who has access to what data across technologies such as Hive, HDFS, and Kafka ("Apache Ranger – Introduction", n.d.). It allows administrators to establish rules so that only permitted users can view or change data. It also keeps track of who has accessed what, which is useful for audits and investigations. Ranger has been deployed and promoted in large-scale settings and where multiple teams want access but not complete control ("Cloudera — The Hybrid Data Company", n.d.). Ranger would be advantageous to the Department if tools like Hive or HDFS are used to store pollution logs or air quality records that only certain staff members should have access to.

4.6.2 Google Cloud IAM

Google Cloud Identity and Access Management (IAM) regulates who has access to various aspects of the Google Cloud platform, such as BigQuery and Cloud Storage. It allows permissions to be assigned depending on user roles, making it simple to grant access solely for what is required("Identity and Access Management documentation — Cloud IAM Documentation", n.d.). Climate TRACE and other public dashboards have implemented IAM to protect sensitive emissions and environmental data ("Climate TRACE", n.d.). For the Department, IAM would help manage access across teams, facilitate PDPA compliance, and decrease the danger of data being accessed or changed without permission.

5 Critical Analysis and Final Tool Selection

This section evaluates the tools proposed earlier and selects the most suitable option for each layer of the big data architecture. The decisions are based on the limitations observed in the Department of Environment's current system; mostly that they rely on manual reporting, disconnected sources, and lack the ability to perform large-scale, real-time analysis. We assumed the Department needs to work with varied data types from sensors, APIs, and field reports, and that they need faster response times, more automation, and better insight for decision-making. While multiple tools were considered, the final selection for each layer reflects the one that best solves the specific challenges, even if it comes with trade-offs.

5.1 Ingestion Layer Selection

NiFi, Kafka, and Flume were proposed. Flume is limited to log data and doesn't support more complex or varied data sources. Kafka is powerful for real-time ingestion and handles high throughput well, but it's harder to set up and maintain, especially for teams without deep infrastructure support. NiFi supports both streaming and batch data, has an easy-to-use interface, and gives control over routing, scheduling, and data flow. One downside is that it's less optimised for extremely high-volume event streams compared to Kafka. Even so, NiFi was selected because it balances functionality with ease of use and fits the Department's need to handle different types of environmental data from multiple sources without requiring deep technical setup.

5.2 Storage Layer Selection

In the storage layer, Hive, Google Cloud Storage, and Amazon S3 were considered. Hive is good for batch data and integrates well with Hadoop-based systems, but it's tied to on-premise infrastructure, which may not suit a more scalable and modular setup. Google Cloud Storage works well, especially if other Google tools are being used, but locks the architecture into a specific ecosystem. S3 is widely used, cloud-based, and integrates with a broad range of processing and analytics tools, including Spark and Trino. A drawback is that deeper integrations might require more setup depending on the tools, but S3 was chosen because of its flexibility, long-term durability, and the fact that it keeps the architecture more open and cloud-agnostic.

5.3 Processing Layer Selection

Spark, Flink, and Dataflow were proposed for this layer. Dataflow is fully managed and easy to use but only runs on Google Cloud, which limits flexibility. Flink is great for real-time event processing, especially with sensor data, but it can be harder to tune and manage. Spark supports both batch and stream processing, has strong community support, and comes with built-in ML and SQL capabilities. It does require some setup and resource tuning, but the trade-off is its flexibility and how well it handles complex transformations. Spark was selected because it can be used for both historical and near real-time data processing, and fits the Department's needs to clean, transform, and prepare large environmental datasets coming from different sources.

5.4 Analysis Layer Selection

Hive, BigQuery, and Trino were proposed. Hive is slower and mainly designed for batch querying. BigQuery is fast and easy to use but only works in Google's ecosystem, which may not fit if the Department wants to stay flexible. Trino allows SQL queries across different systems like Hive, S3, and relational databases without moving data. One of its limitations is that it is

not built for heavy processing, only querying. But Trino was selected because the Department will likely have data spread across multiple sources, and Trino makes it possible to explore that data without having to centralise it all, which saves time and effort in the long run.

5.5 Visualisation Layer Selection

Power BI and Tableau were considered. Power BI is easy to use and works well for reports but is more tailored to Microsoft environments and has weaker geospatial support. Tableau supports strong map-based visuals, works with various data types, and is commonly used for public dashboards in environmental projects. It does have licensing costs and a slightly steeper learning curve, but Tableau was selected because its flexibility, visual richness, and support for spatial layers make it a better fit for showing things like air quality zones, rainfall maps, or emissions by region.

5.6 Security and Governance Selection

Apache Ranger and Google Cloud IAM were proposed. IAM is easy to use within Google Cloud and gives good role-based access control, but doesn't apply outside that ecosystem. Ranger supports fine-grained access control and auditing across Hadoop, Hive, and Spark. It can take more effort to set up, but Ranger was chosen because it fits better with the rest of the tools selected especially Spark, Hive (if still partially in use), and NiFi.It also supports detailed policy management and audit logs, which the Department will need for internal controls and future compliance.

5.7 Final Architecture Overview

This architecture was built by choosing the best tool at each layer based on functionality, not vendor alignment. The result is a modular, open architecture that doesn't lock the Department into any one cloud provider. NiFi handles varied data sources smoothly, S3 offers flexible storage, Spark enables scalable processing, Trino supports cross-platform analysis, Tableau brings strong visualisation, and Ranger ensures secure access and auditing. This setup solves the main gaps identified in the case study like slow manual reporting, fragmented systems, and limited analytics and enables the Department to carry out meaningful big data analytics for smarter, faster environmental decision-making.

6 Security and Privacy Considerations

When dealing with large volumes of environmental data, there are a few privacy and security issues that need to be considered early on. One of the main concerns is unauthorised access. If different teams or departments are using the system, there's always the chance that someone

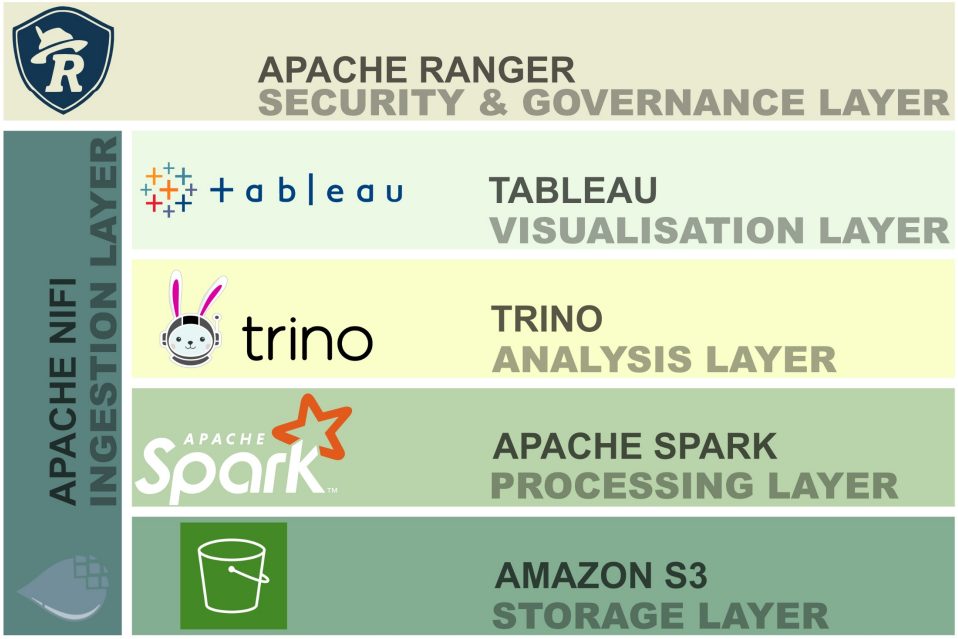

Figure 2: Proposed Big Data Ecosystem

might access data they are not supposed to. This is especially important if the data includes location-based pollution levels or citizen-submitted reports. Without proper access controls, there's a risk of accidental changes or misuse that could affect reporting accuracy or public trust.

Another issue is data being exposed while it is moving through the system. For example, sensor data or file uploads might pass through insecure connections if encryption is not applied properly. This could make the system vulnerable to tampering or data leaks, which would go against laws like the Malaysian PDPA. It's also important to make sure sensitive data is protected when stored, especially if it includes personal or location-specific information.

Furthermore, not having clear logs or tracking is another risk. If someone makes a mistake or accesses something they shouldn't, the Department needs a way to trace back what happened. Without audit trails, it becomes hard to investigate issues or hold users accountable.

To reduce these risks, several tools in the architecture already support strong security features. Apache Ranger can manage who is allowed to access specific files or data in tools like Hive or Spark, and it logs every access request for review. Amazon S3 allows permissions to be set for different folders or buckets and supports encryption for files stored on the cloud. Tableau also supports role-based access, so that only authorised users can view certain dashboards or reports On top of that, the whole system should use secure connections and encryption between services, including during ingestion through NiFi, to make sure no data gets intercepted. So while these risks are real, they can be managed well if the right access controls, encryption, and monitoring tools are used consistently across the system.

Bibliography

Aclima case study. (n.d.). *Google Cloud*. https://cloud.google.com/customers/aclima

Apache flink documentation. (2024). *Apache.org*. Retrieved May 21, 2025, from https://nightlies.apache.org/flink/flink-docs-release-1.17/

Apache Hadoop. (n.d.). Apache hadoop. *hadoop.apache.org*. Retrieved May 21, 2025, from https://hadoop.apache.org

Apache kafka. (n.d.). *Apache Kafka*. Retrieved May 21, 2025, from https://kafka.apache.org/documentation/

Apache nifi user guide. (n.d.). *nifi.apache.org*. Retrieved May 21, 2025, from https://nifi.apache.org/docs/nifi-docs/html/user-guide.html

Apache ranger – introduction. (n.d.). *ranger.apache.org*. https://ranger.apache.org

Apache Spark. (2019). Overview - spark 2.4.4 documentation. *Apache.org*. Retrieved May 21, 2025, from https://spark.apache.org/docs/latest/

AWS. (2022). Amazon simple storage service documentation. *Amazon.com*. Retrieved May 21, 2025, from https://docs.aws.amazon.com/s3/

Bigquery documentation. (n.d.). *Google Cloud*. Retrieved May 21, 2025, from https://cloud.google.com/bigquery/docs

BU, Coforge-Salesforce. (2017, February). Bright lights, smart city, big data. *Coforge.com*. Retrieved May 21, 2025, from https://www.coforge.com/what-we-know/blog/bright-lights-smart-city-big-data

Climate trace. (n.d.). *climatetrace.org*. https://climatetrace.org

Cloud storage documentation. (n.d.). *Google Cloud*. Retrieved May 21, 2025, from https://cloud.google.com/storage/docs

Cloudera — the hybrid data company. (n.d.). *Cloudera*. https://www.cloudera.com

Composting - solid waste management facilities map. (n.d.). *State of New York*. https://data.ny.gov/Energy-Environment/Composting-Solid-Waste-Management-Facilities-Map/y4ic-kfv6

Data exploration with tableflow, apache iceberg, and trino. (2025, April). *Confluent*. Retrieved May 21, 2025, from https://www.confluent.io/blog/building-streaming-data-pipelines-part-1/

Dataflow documentation. (n.d.). *Google Cloud*. Retrieved May 21, 2025, from https://cloud.google.com/dataflow/docs

DEFRA. (2019). Data archive- defra, uk. *Defra.gov.uk*. https://uk-air.defra.gov.uk/data/

Dinakar, R. (2024). *Real-time iot sensor data streaming and processing with apache flink: A scalable solution for smart monitoring* (tech. rep.).

Flume 1.11.0 user guide — apache flume. (n.d.). *flume.apache.org*. Retrieved May 21, 2025, from https://flume.apache.org/FlumeUserGuide.html

Greca, S., Shehi, I., & Nuhi, J. (2023, April). *Analyzing climate changes impacts using big data hadoop* (tech. rep.). https://ceur-ws.org/Vol-3402/paper03.pdf

Home - apache hive - apache software foundation. (n.d.). *cwiki.apache.org*. Retrieved May 21, 2025, from https://cwiki.apache.org/confluence/display/Hive/Home

Identity and access management documentation — cloud iam documentation. (n.d.). *Google Cloud*. https://cloud.google.com/iam/docs

Kenyon, O. (2017). Using kafka and grafana to monitor meteorological conditions. *Scott Logic*. Retrieved May 21, 2025, from https://blog.scottlogic.com/2017/10/13/MetOfficeKafka.html

Markert, K. N., da Silva, G., Ames, D. P., Maghami, I., Williams, G. P., Nelson, E. J., Halgren, J., Patel, A., Santos, A., & Ames, M. J. (2024). Design and implementation of a bigquery dataset and application programmer interface (api) for the u.s. national water model. *Environmental Modelling and Software*, *179*. https://doi.org/10.1016/j.envsoft.2024.106123

Microsoft industry clouds. (n.d.). *www.microsoft.com*. https://www.microsoft.com/en-us/industry?post_type=articles&category=&product=microsoft-power-bi

Racka, K., & Płocku, M. U. P. (2022). Apache nifi as a tool for stream processing of measurement data summary. https://doi.org/10.19251/ne/2022.35(7)

Shih, D. H., To, T. H., Nguyen, L. S. P., Wu, T. W., & You, W. T. (2021). Design of a spark big data framework for pm2.5 air pollution forecasting. *International Journal of Environmental Research and Public Health*, *18*. https://doi.org/10.3390/ijerph18137087

Tableau. (2023). What is tableau? *Tableau*. https://www.tableau.com/why-tableau/what-is-tableau

Trino documentation — trino 475 documentation. (2025). *Trino.io*. Retrieved May 21, 2025, from https://trino.io/docs/current/