EMPLOYABILITY OF DATA MODELING AND ANALYTICAL TOOLS AND TECHNIQUES FOR EFFECTIVE MANAGEMENT OF BIG DATA IN THE ENERGY DATA PLATFORM

Chandan Bansal

Shivaji College, University of Delhi, Delhi, India

ABSTRACT

The energy sector encompasses many industries, including oil and gas, coal, minerals, renewable energy, and electricity. Managing this domain's vast and diverse datasets presents challenges regarding volume, variety, veracity, and velocity. This work focuses on developing a robust data model tailored to address these challenges within the energy sector. By analyzing the national energy significant data architecture, the research aims to create data mapping and correlation techniques, implement extensive data master data management utilizing energy industry standardization, and design a big data management portal for upstream and downstream energy activities. This research aims to lay the groundwork for technology and big data management solutions within the energy sector, encompassing petroleum, coal, geothermal, and renewable energy. Furthermore, it seeks to pave the way for predictive analytics and optimization of national energy production processes in the future.

INTRODUCTION

Fossil energy resources like oil and gas, coal, minerals, and new and renewable energy sources remain pivotal in supporting national economies and energy demands. Traditional methods of managing these resources through exploration and production now require modernization with information, data, and technical analysis. A crucial aspect of this modernization is establishing a data management system capable of predicting energy consumption and production, a cornerstone for achieving developmental milestones. Big data technology offers avenues for efficient data collection, orchestration, mapping, and analysis, thereby facilitating the transformation of national energy landscapes, particularly in Industry 4.0.

Data exhibits diverse characteristics in terms of volume, types, variations, and interpretations, especially within the expansive scope of energy data, encompassing oil and gas, coal, minerals, renewable energy, electricity, and more. This data spans the entire energy lifecycle, from exploration and production to downstream activities like consumption, distribution, and utilization. Effective energy data management within an extensive national data management system promises significant advantages, providing opportunities for resource control and maximizing community benefits.

This research endeavours to develop technology and big data management solutions capable of predictive analysis and enhancing national energy production. The primary focus lies in establishing a data model for energy data management to support implementing a comprehensive, extensive data energy management system.

The outcomes of this work include the development of a data model and implementation strategies for the Big Energy Data Platform. These achievements stem from analyzing national energy significant data architecture, developing master data management standards within the energy industry, and crafting tailored big data management solutions for upstream and downstream energy operations. Ultimately, this research lays a foundational framework, particularly for extensive data management in the energy sector, encompassing petroleum, coal, geothermal, and renewable energy, with the potential for predictive analysis and optimization of national energy production processes.

DATA MODELLING AND BIG DATA

Developing extensive data management systems involves various components such as data centres, management systems, extensive data systems, data analysis, and utilization. Numerous research efforts have contributed to advancing these components, as evidenced by projects like FutureGrid, DIET, BEinGRID, ScienceForge, DALA Project, and GamaCloud.

FutureGrid is a testbed for Grid and Cloud Computing, integrating multiple cloud infrastructures and researching authentication, authorization, scheduling, virtualization, and cloud-based computing. DIET, introduced by INRIA in 2000, implements distributed scheduling on grid and cloud infrastructure, emphasizing demand-based resource allocation and cloud economics. BEinGrid provides a grid infrastructure for real business scenarios, addressing cost reduction, performance improvement, business model development, and Software as a Service (SaaS) implementation.

ScienceForge is dedicated to developing cluster infrastructure for collaborative research data, utilizing an Application Framework to offer SaaS services for data collection, processing, and archiving. The DALA Project, a continuation of ScienceForge, focuses on data preservation using cluster infrastructure, resulting in a comprehensive data management model encompassing input, memory, preservation, output, management, and archival storage layers. GamaCloud implements a scientific research infrastructure based on the Grid model, managing fabric components, network management, and cluster storage, leveraging the DALA project's Globus Middleware services and pre-built data management services.

Organizations like The McKinsey Global Institute define Big Data as datasets exceeding traditional database software capacity for management and analysis, arising from data transactions, interactions, and continuous observation. Big Data is characterized by its immense volume, variety, velocity, and veracity. Dodson emphasizes characteristics such as large data volumes with varied formats, types, and structures generated in real-time. The 4V or 5V model encapsulates these characteristics, emphasizing Volume, Velocity, Variety, Veracity, and sometimes Value.

Viana et al. proposed a reference architecture for archiving, preservation, and retrieval in the Big Data environment, aiming to address the absence of a specialized architecture for structured and unstructured data preservation. Their research aims to provide a set of patterns for concrete architectures supported by artefacts for real-world implementation.

Nguyen et al. present a Content Server system architecture designed to address challenges in Large-scale Digital Preservation Archive Systems (LDPAS), focusing on flexibility, scalability, configurability, and service orientation. Their Content Server design includes storage and search servers, Hierarchical Storage Management System deployment for digital asset storage, and a metadata repository based on XML database clusters for enhanced performance and availability.

DATA STANDARD IDENTIFICATION

This phase involves the architectural design process, encompassing the design of the extensive data energy infrastructure, compilation of technology components constituting the architecture, compilation of significant data management components, preparation of process and analytical significant data components, and creation of user interfaces in mockup form. The outcome of this design phase is the development of a prototype for the considerable data energy architecture, comprising architecture, data management, metadata, and identifiable data sources. Dissemination activities are also conducted during this phase to gather comprehensive feedback.

The next step involves developing Master Big Data with industry standardization. This entails aligning extensive data development with industry standards to ensure compliance with international norms. Technical analysis is performed, and the design outcomes from the previous stage are adjusted accordingly, mapping them with global standards about energy data. The chosen standard for this endeavour is the Professional Petroleum Data Management standard. This stage includes workshops, expert discussions, feedback workshops, and expert reviews from the industry to obtain thorough and in-depth insights.

Identification of energy data management standards globally is essential, including:

- The Open Group Open Subsurface Data Universe™ (OSDU) Forum
- International Association of Oil & Gas Producers (IOGP)
- POSC Caesar Association
- Petroleum Industry Data eXchange (PIDX)
- Pipeline Open Data Standard (PODS)
- PPDM Association
- Open Geospatial Consortium (OGC)
- Object Management Group (OMG)
- Pipeline Open Data Standard (PODS)
- PPDM Association
- Open Geospatial Consortium (OGC)
- Object Management Group (OMG)

ENERGY DATA MODEL AND IMPLEMENTATIONS

The efforts have yielded the development of the data model and implementation strategies for ample data storage. The data model results from analyzing the national energy significant data architecture, incorporating extensive data master development management aligned with energy industry standardization, and addressing upstream and downstream energy management.

A. Data Models and Identification

Data models are crafted by identifying data attributes defined by various standards. These attributes are structured into nodes, sub-nodes, and leaf nodes, representing energy data categories, attributes, and values. The process involves referencing standards such as Professional Petroleum Data Management and engaging in workshops, expert discussions, and feedback sessions to ensure comprehensive review and alignment with industry practices.

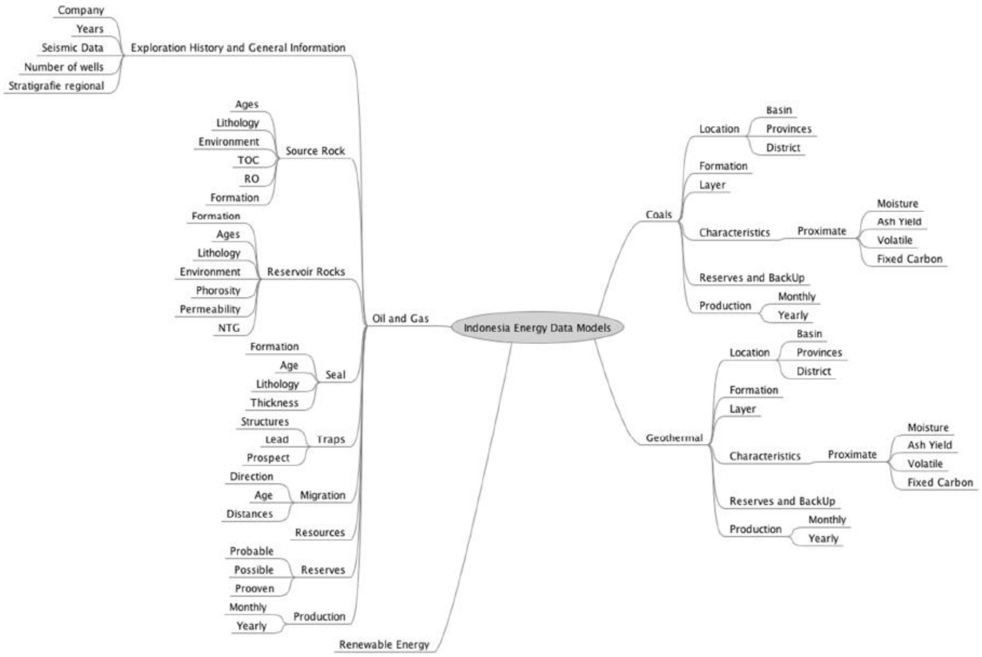

Fig. 1. Data Model Maps & Identification

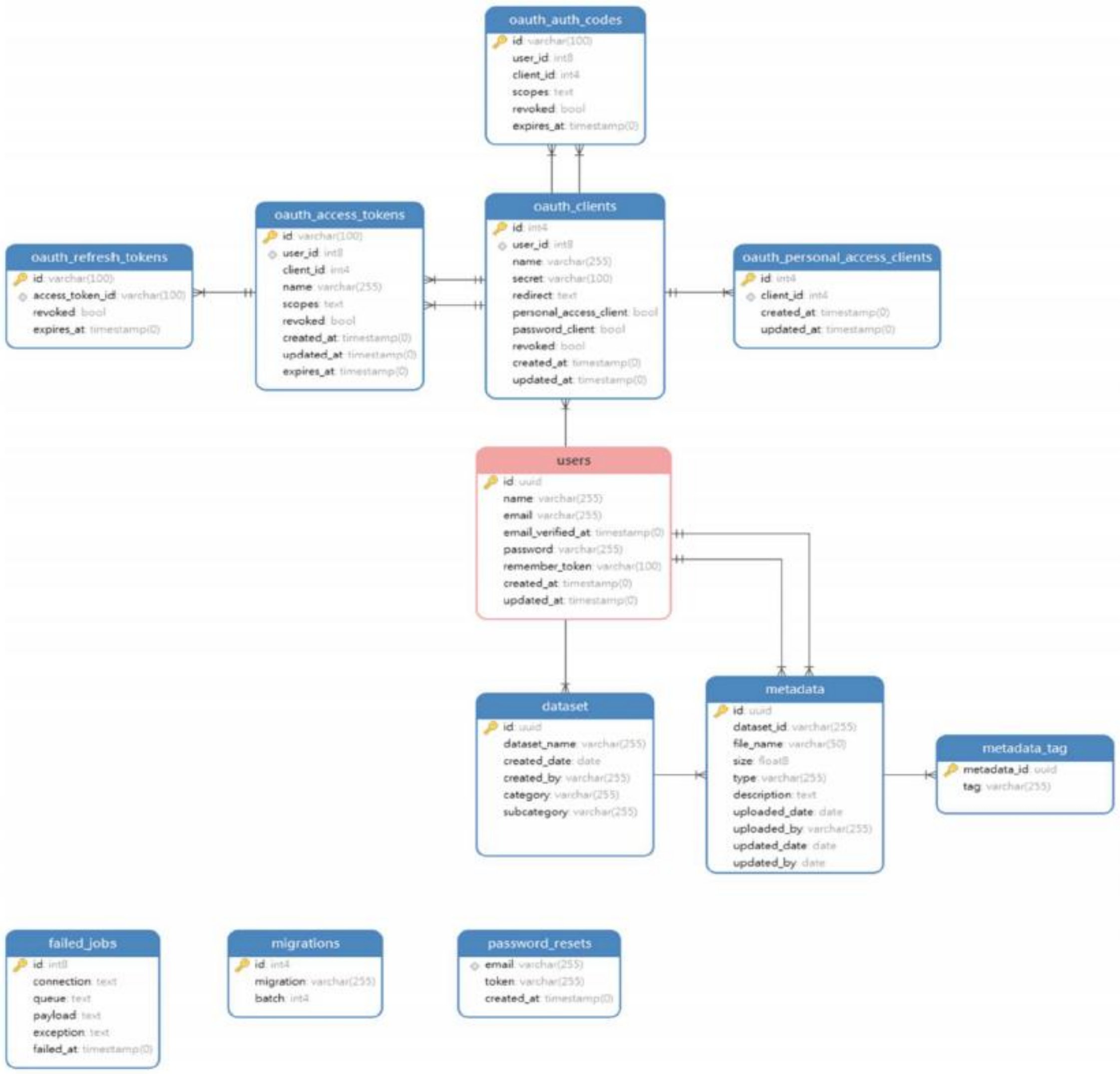

Fig. 2. Big Data Storage Scheme Implementation

B. Big Data Storage Implementation

For storage, PostgreSQL and MongoDB version 4.4 databases are utilized. PostgreSQL manages user data relations and file metadata with an additional scheme for storing user data from Firebase. MongoDB is explicitly chosen for its compatibility with the LAS file viewer library, which requires JSON-form storage. This dual database approach ensures efficient management of structured and unstructured data, catering to diverse storage needs within the context of energy data.

CONCLUSION AND DISCUSSIONS

Energy data crucial for national production are managed by various stakeholders, including government entities, industries, professionals, and institutions. These data exhibit diverse characteristics in terms of volume, types, variations, and interpretations, spanning a broad spectrum of energy sources such as oil and gas, coal, minerals, new and renewable energy, conversion energy, electricity, and more. They encompass data from upstream processes like exploration, production, and maintenance and downstream activities, including consumption, distribution, and utilization.

The outcomes of this work include the development of a data model and implementation strategies for the Big Energy Data Platform. This involved analyzing the national energy significant data architecture, establishing extensive data master development management aligned with energy industry standards, and implementing big data management solutions for upstream and downstream energy operations. The research has laid a foundational framework for extensive data management in the energy sector, covering petroleum, coal, geothermal, and renewable energy, facilitating predictive analysis and national energy production optimization.

The results have been implemented in the form of the big data energy portal "GamaBox Explorer," as depicted in Figure 3. This portal is a comprehensive platform for accessing and analyzing energy data, further enhancing the management and utilization of energy resources for national development.

REFERENCES

[1] Teng, F., Management Des Donnees Et Ordinnnancement Des Taches Sur Architectures Distributes, Desertation, Ecole Cenrale Paris Et Manufactures, Centrale Paris 2012.

[2] Riasetiawan, M., Dala Project: Digital Archive System for Long Term Access. The Second International Conference on Distributed Framework and Applications (DfmA) 2010, 2-3 Agustus 2010 IEEE, pp. 1-5, FMIPA UGM, Yogyakarta Indonesia.

[3] Riasetiawan, M., CloudBox: a cloud technology on the box. 8th e- Indonesia Initiative Forum 2012.

[4] Riasetiawan, M., GamaCloud: The Development of Cluster and Grid Models based Shared memory and MPI, CITEE 2012, Yogyakarta Indonesia.

[5] Riasetiawan, M., Mahmood, A.K, Science-Forge: A collaborative scientific framework, 2010 IEEE Symposium on Industrial Electronics & Applications (ISIEA), Penang Malaysia, 3-5 Oktober 2010, pp.665- 668, DOI: 10.1109/ISIEA.2010.5679381.

[6] Riasetiawan, M., Mahmood, A.K., Managing and Preserving Large Data Volume in Data Grid Environment, 2010 International Conference on Information Retrieval and Knowledge Management (CAMP'10), 17-18 Maret 2010 IEEE, pp. 91-96, Shah Alam Selangor Malaysia.

[7] Mayinka, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., 2011, Big Data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute 2011 Report, [Online], Mei 2011, available at http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

[8] Dodson, V., 2013, Big Data and Business Intelligence, eCLIPSeCon, Boston 2013, http://www.eclipsecon.org/2013/sites/eclipsecon.org.2013/files/BigData_and_BusinessIntelligenc e_EclipseCon2013.pptx.

[9] Viana, P, Sato, L. References Architecture for Long-term Archiving, Preservation, and Retrieval of Big Data, International Conference on Trust Security and Privacy in Computing and Communication, IEEE Conference Publication, 2014