1. GraphFrames and Databricks: Technologies for Data Analysis in Energinet

Introduction: Energinet faces increasing complexity in the energy system as conventional power plants are phased out and renewable sources integrated. Data and digitalization have become **crucial catalysts** for handling core tasks efficiently and ensuring a robust future energy system[1]. To support goals of **cost savings**, **green transition**, and **resilience**, Energinet must leverage modern data analytics tools. Two promising technologies in this context are **GraphFrames** and **Databricks**. This report examines what GraphFrames and Databricks are from a technical standpoint, and how they can be applied in Energinet to strengthen decision-making, increase efficiency, and drive innovation.

Data as a Key Enabler

Energinet is harnessing the vast potential of data and digitalization as a *critical catalyst* for efficient operations and a robust future energy system.

Graph Analytics & Big Data

GraphFrames enables large-scale network analysis on big data, while **Databricks** provides a scalable platform for general big data processing and AI.

---

2. GraphFrames – Graph Analytics with Apache Spark

What is GraphFrames? GraphFrames is a library for Apache Spark that extends Spark's DataFrame API with graph-processing capabilities[2]. It allows data to be represented as graphs (nodes and edges) and performs *graph algorithms* at scale using Spark clusters. GraphFrames can be seen as an evolution of Spark's built-in GraphX library. Whereas GraphX is primarily Scala-based, GraphFrames provides a more user-friendly DataFrame-based interface and supports Python and SQL. This makes it easier for data science teams to work with graph data without leaving the familiar Spark ecosystem.

Technical details and capabilities: GraphFrames offers a range of built-in algorithms and tools for graph analysis:

- **Standard graph algorithms:** The library includes around *48 pre-built graph algorithms*, such as PageRank, Breadth-First Search (BFS), Connected Components, triangle count, and others[2]. These cover classic network analyses and even some "graph machine learning" methods (e.g. node embeddings for ML)[2]. The algorithms execute in parallel across the cluster, enabling analysis of very large graphs.
- **Motif finding (pattern search):** GraphFrames introduces a *motif-finding DSL*, a declarative way to search for patterns in the graph[2]. For example, one can query for all occurrences of a certain connection pattern (subgraph) in the grid network graph. This is useful for detecting relationships or structures of interest (e.g. a cyclic connection between substations indicating redundancy).
- **Integration with DataFrames:** Because GraphFrames extends Spark DataFrames, it's easy to combine graph operations with regular data processing. For instance, you can filter nodes/edges with SQL-like syntax before running a graph computation. Results from graph algorithms (such as a centrality score per node) become available as DataFrames that can be joined with other datasets. This integration provides flexibility to blend graph insights with other data (measurements, financial data, etc.) in the same analysis workflow.

Built-in Graph Algorithms

48+

Standard algorithms (e.g. PageRank, component analysis) available out-of-the-box in GraphFrames.

- **Scalability:** By running on Apache Spark, GraphFrames is horizontally scalable. Large graphs are partitioned and processed in memory across many machines in the Spark cluster. This means graph analysis is not limited by single-machine memory – a crucial point since Energinet may work with graphs comprising thousands of network nodes and connections. The Spark engine handles distribution of data and computations, ensuring GraphFrames can utilize cluster resources efficiently[2].
- **Databricks compatibility:** GraphFrames is compatible with the Databricks platform, since Databricks is built on Apache Spark. In fact, an internal colleague noted that because many teams and departments already use Databricks, it makes sense to explore GraphFrames for graph analytics on that platform[3]. In other words, Energinet can leverage existing Databricks infrastructure for graph analysis without introducing a new separate system – making adoption of GraphFrames relatively *cost-light* and accessible.

Example application: Imagine Energinet's electricity transmission network represented as a graph, where nodes are transformer stations and edges are high-voltage connections. With GraphFrames, one could run a **PageRank** algorithm to identify the most *central substations* in the network (i.e. those most critical for power flow)[2]. Likewise, a *Connected Components* algorithm could reveal any isolated "islands" in the network under certain failure scenarios. GraphFrames' motif search could find structural patterns – for example, detecting if "triangle" loops exist in the network topology that provide redundancy against outages. These analyses are performed on the entire grid graph at once using parallel computation, which would have been very time-consuming before. The result is concrete insight into the network's structure and vulnerabilities based on data-driven analysis.

3. Databricks – Platform for Big Data and Machine Learning

What is Databricks? Databricks is a *commercial data analytics platform* in the cloud, built on Apache Spark. Databricks adds an easy-to-use layer on top of Spark and manages much of the infrastructure, so users can focus on data and analysis rather than setup. In short, **Databricks provides a fully managed Spark environment** along with extra features[4]:

- **Managed Spark clusters:** Databricks takes care of provisioning and managing Spark clusters. This means developers do not need to manually configure master/worker nodes, YARN, or Spark versions. Clusters can automatically scale up and down as needed (*auto-scaling*), and data processing is distributed efficiently across the cluster[4]. The platform allocates resources optimally, which **frees up time** and reduces the need for deep expertise in cluster administration.
- **Collaborative workspace:** One of Databricks' most visible features is its **notebook interface**. In a shared workspace, multiple users can develop code (in Python, R, Scala or SQL) together, visualize results, and document insights side by side. It's akin to an advanced Jupyter Notebook running on Spark in the cloud. This collaboration interface promotes knowledge sharing and makes it easier for data engineers, analysts, and data scientists to work together on projects. Code and results are versioned and shared in real time, reducing silos between teams.
- **Data and storage integration:** Databricks integrates with data lakes, data warehouses, and other sources. The platform supports direct access to, for example, Azure Data Lake Storage, SQL databases, and streaming sources. Additionally, Databricks includes **Delta Lake**, a technology for reliable data lakes with ACID transactions. For Energinet, this means raw SCADA data, historical measurements, GIS information, etc., can be brought together in one place and analyzed in a unified way. Databricks can serve as a single "*source-of-truth*" for data – indeed a stated goal of Energinet's data platform is

to create one single source-of-truth and enable transparent, secure analysis across the business[4].

- **Advanced analytics and ML:** On top of the Spark engine, Databricks offers built-in capabilities for machine learning and real-time streaming. It is easy to scale from batch data engineering to real-time streaming analysis within the same platform. Moreover, MLflow is supported to track machine learning experiments and models. For Energinet, this could be used to develop and deploy forecasting models for electricity load or wind production at scale – all on the same platform.

User-Friendly Spark

Databricks provides a **collaborative workspace** with notebooks, making it easy for multiple users to work in parallel on analyses and see results immediately.

Auto-Managed Clusters

The platform handles **auto-scaling of clusters** and infrastructure, so developers can focus on data rather than ops setup.

Unified Data Access

Databricks integrates with data lakes and databases, providing a single source-of-truth for data and enabling unified analysis across the business.

Databricks in Energinet: The platform is already widely used internally – many teams and departments at Energinet utilize Databricks as their primary data tool[3]. This suggests the organization has embraced cloud-based big data analytics and that there is a common environment where data can be shared and analyzed. Databricks aligns well with Energinet's strategy of intensifying digitalization: by centralizing data handling and analysis, the platform enables users to **identify patterns and trends in data, make better decisions, and optimize processes and services** across business areas[4]. At the same time, it ensures scalability and future-proofing, as the platform can grow with the task and accommodate new data sources or analysis types as needs evolve.

To summarize the characteristics of the two technologies, below is a comparison:

Feature	GraphFrames	Databricks
Type	Spark library for graph analytics (open source)[2]	Managed big data/AI platform (commercial)[4]
Primary purpose	Large-scale graph algorithms and pattern search[2]	Efficient big data processing and collaborative analytics[4]
Key features	~48 built-in graph algorithms; motif query DSL[2]	Notebook interface, data lake integration, auto-scaling clusters[4][4]
Scalability	Runs on Apache Spark clusters (horizontal scale-out)[2]	Fully managed auto-scalable clusters in the cloud[4]

(Table 1: Overview of GraphFrames vs. Databricks)

---

4. Potential for Energinet: Cost Savings, Green Transition, and Resilience

With an understanding of GraphFrames and Databricks, let us consider how they can **benefit Energinet** in terms of economics, the green transition, and system robustness. The combination of advanced graph analytics and a strong data platform can support initiatives that address Energinet's future challenges:

Greater Efficiency & Savings

Data insights from the Spark platform can optimize operations and maintenance, cutting costs without compromising reliability.

Data-Driven Green Transition

Advanced analytics help integrate more renewables by accurately forecasting production/consumption and optimally planning the grid.

Enhanced Resilience

Graph analysis identifies critical nodes and vulnerabilities in the energy network, while real-time data monitoring enables faster response to events.

5. Increased efficiency and cost savings

A core part of Energinet's mission is to ensure *economic efficiency* in the energy sector – delivering green energy and high security of supply at a cost society can bear[5]. GraphFrames and Databricks can contribute to this goal by optimizing operations and decision processes through data:

- **Optimizing grid investments:** Using graph analysis, Energinet can better assess where to invest in grid improvements. GraphFrames can reveal which connections or stations are most *critical to network connectivity* (e.g. via centrality measures). Investments can be targeted to reinforce the weakest links in the system first, yielding the most "security of supply per krone." Conversely, graph analysis might highlight redundant structures where capacity could be rationalized, avoiding over-investment. This helps ensure both under- and over-investment are avoided, directing capital efficiently.
- **Improved asset management and maintenance:** Maintenance plans can be optimized by combining GraphFrames analysis with equipment telemetry in Databricks. For example, Energinet has sensor data from transformers (temperature, load, vibrations, etc.) in its data platform. By coupling this with graph information about the network structure, one can pinpoint *critical components* that are both centrally located in the graph **and** showing signs of wear. These assets would be high priority for predictive maintenance. Data-driven models can predict failures before they occur, saving costs by preventing expensive outages. All this requires handling large amounts of data – a task Databricks is designed to handle. Indeed, Energinet's digitalization agenda aims to **optimize task execution** and create efficiency across the organization by exploiting data[4]; GraphFrames and Databricks are concrete tools to achieve these optimizations.
- **Efficient operations and market balancing:** Databricks can be used to analyze huge volumes of operational data to find inefficiencies. For instance, one could analyze historical balancing market data and consumption forecasts on the platform to identify patterns where too much reserve is consistently procured, or where forecasts systematically err. Using machine learning on Databricks, forecasts (e.g. for demand or renewable output) can be made more accurate, so Energinet procures exactly the reserve needed – no more, no less. This directly **reduces operating costs**. An internal initiative already points in this direction: Energinet wants to exploit data and digitalization to **optimize task execution** and create *efficiencies* across the organization[4]. GraphFrames and Databricks are concrete tools that can realize such efficiencies.

6. Data-driven green transition

Denmark has ambitious goals for a fully green energy sector, and Energinet must enable integration of large amounts of solar and wind energy. Both GraphFrames and Databricks can support this **green transition**:

- **Planning renewable integration:** GraphFrames can model the electric transmission network as a graph and test different scenarios for connecting new renewable sources. By adding new nodes (e.g. wind farms or energy islands) and edges (cables) to the graph, and running graph algorithms, planners can quickly evaluate the impact on network structure. For example, a community detection algorithm could run on the graph to see if new wind parks create standalone "islands" in the topology that should be better connected, or if they integrate smoothly. GraphFrames can also help with **capacity analysis**: a BFS or shortest-path search might expose bottlenecks in the paths between major generation sites and load centers, indicating where grid expansion is needed to avoid future overloads.
- **Improved generation and load forecasting:** Databricks' strong ML capabilities can be used to build more advanced prediction models for renewable production and electricity consumption. The platform can handle decades of historical data combined with weather forecasts and real-time sensor feeds. This is essential for the green transition, since accurate forecasts reduce the need for standby reserves and allow the grid to operate closer to its limits with high renewable penetration. Better forecasts mean less renewable energy is wasted (fewer instances of curtailing wind turbines), which in turn allows a higher share of green energy in the system. Energinet is already looking at AI-based weather models to improve, for example, wind forecasts[6] – such a solution could potentially run on a platform like Databricks to easily scale with data. In fact, studies have shown that advanced AI weather models like **GraphCast** can improve wind forecasting accuracy by ~15%, leading to substantial cost savings in balancing markets[6]. By deploying similar models, Energinet can reduce reserve requirements and operational costs while integrating more renewables.
- **Holistic energy system analysis:** The green transition involves sector coupling (e.g. power-to-X, electrification of transport, etc.), making the energy system even more complex. Databricks allows analyzing cross-domain data – electricity, gas, district heating, consumption patterns – together. GraphFrames could be used to build *multi-layer graphs* modeling different parts of the energy system and their interconnections. By leveraging these tools in tandem, Energinet can identify new opportunities for synergy and efficient use of green energy that might not be discovered by looking at each part in isolation.

7. Strengthened resilience and security of supply

Security of supply – or resilience – is the ability to withstand and cope with disruptions in the energy system. GraphFrames and Databricks can help Energinet maintain Denmark's high security of supply, even in an era of major system changes:

- **Identifying critical nodes:** Through graph analysis (GraphFrames), one can perform *vulnerability analyses* on the network model. For example, remove a node or edge in the graph (simulate an outage) and use algorithms to determine if the graph splits into disconnected components or remains connected. By iterating over all possible outages, GraphFrames can highlight those elements where a single failure would fragment the system (typically very high *betweenness centrality* nodes). These points are obvious candidates for grid reinforcement or close monitoring. Energinet can thus prioritize adding redundancy exactly where the graph model shows it is thinnest – making the entire grid more robust to failures. Such a data-driven approach complements classic engineering judgment with quantitative evidence.
- **Faster fault localization:** In operations, GraphFrames can also assist in real-time or near-real-time understanding of network problems. If a fault occurs (e.g. a power outage in a region), a graph representation can immediately show which parts of the

network are **downstream** and impacted. GraphFrames can rapidly compute, for instance, the *connected components* of the remaining graph, thereby identifying which areas are cut off. This information could be integrated into control center tools to give an instant picture of the situation, based on data.

- **Monitoring and anomaly detection:** Databricks can handle streaming data from sensors and SCADA systems, opening the door for *anomaly detection* in real time. By drawing on historical patterns (learned by ML models on the platform), the system can flag unusual behavior – e.g. a transformer suddenly showing strange voltage fluctuations – which may be early signs of an impending issue. This enhanced **situational awareness** is critical for resilience, as it provides time to respond preventively. If combined with graph context, the system would know not only *that* a given sensor reading is abnormal, but also *where* in the network and which other components would be affected if something goes wrong there.
- **Simulation and emergency drills:** The data-rich platform can be used to run what-if simulations. For example: what if a major transmission line to another country trips during peak load? Using historical load and generation data in Databricks plus the graph model for network structure, one can simulate the consequences and test if contingency plans and backup capacity are sufficient. Such a simulated *digital twin* of the power grid – driven by real data and graphs – enables improving emergency plans and ensuring Energinet can withstand even rare but critical events.

---

8. Conclusion

Closing remarks: GraphFrames and Databricks represent two complementary technologies that both address the need for **data-driven insight** at Energinet. GraphFrames provides the means to understand complex relationships in the energy system through graph theory and large-scale analysis, while the Databricks platform delivers the necessary infrastructure to collect, process, and exploit massive data across the organization.

Energinet's strategy points towards a single data platform and leveraging data as a central part of future solutions[4]. In light of this, these two technologies are an excellent fit: Databricks is *becoming* the common data platform for many teams[3], and GraphFrames can be incorporated to add specialized graph analytics without requiring new standalone systems. By applying these tools in the right areas, Energinet can achieve:

- **Data-driven savings:** through optimized operations, better investment prioritization, and reduced waste of resources.
- **Acceleration of the green transition:** via smarter grid planning and more accurate forecasts, enabling higher shares of renewables.
- **Higher resilience:** by proactively identifying and mitigating vulnerabilities in the energy system, and responding faster to incidents.

In summary, GraphFrames and Databricks can **make Energinet better prepared** for the future. As was highlighted internally, it is sensible to take advantage of the fact that *"many teams and departments use Databricks"* already and extend it with GraphFrames to get even more value from our data[3]. By embracing these technologies, Energinet bolsters its position as a data-driven organization – ready to deliver on the energy trilemma: green energy, high supply security, and economic efficiency, all at once[5]. [5]

References

[1] Data i Energinet

[2] Research Graph Database Alternatives

[3] BAD: Daily Standup

[4] Dataplatformen i Energinet
[5] Energinets Leverancemodel
[6] Tapestry for Energinet's Renewables- Formatted