Main Reasons Why Data-Driven Projects Fail and How the Cloud Comes to the Rescue

6 November 2024 - 5 min. read

AI | Cloud adoption | Data and Analytics | Data Ingestion | Data Security and Governance | Machine Learning

Introduction

In the world of innovation, the concept of *fail-fast* refers to the ability to identify and resolve issues quickly, minimizing investment in approaches that may not lead to the desired results. This mindset, increasingly essential in technology projects, helps save time and resources, allowing hypotheses and strategies to be tested in an agile manner. In data-driven projects, where value can only be extracted through iterations and continuous optimizations, the *fail-fast* philosophy provides a method for experimenting, learning, and improving without compromising the entire project.

A data-driven project should not be seen solely in terms of data analysis or processing data using Machine Learning algorithms or implementing Generative AI. The stages of a data-driven project start from the planning and managing of data flow from the origin (such as raw data sources) to final processing and analysis. Generally, we can group the phases of such projects into three main blocks: data ingestion, data management, and data value. At each stage, some pitfalls can lead entire projects to fail. Supporting a well-defined strategy, the Cloud helps maximize the value of data and reduces the risks of errors or inefficiencies that can lead to failure.

Therefore, the Cloud is the perfect enabler of this approach, thanks to its flexibility, high scalability, and innovative services. Unlike on-premises infrastructures, the Cloud allows for the launch of Proof of Concepts (PoC) at reduced initial costs and without the burden of significant infrastructure investments. This enables companies to test quickly and at a low

cost, reducing the financial impact of potential errors in the various stages of data-centric projects.

How the Cloud Improves Data Quality and Availability

Data quality and availability are essential for the success of data-driven projects. A project that begins with poor-quality data or limited access risks failing to deliver meaningful results, incurring economic impacts from development costs. In the Data Ingestion phase, the Cloud, through its managed services and out-of-the-box automation capabilities, helps improve data quality and availability while simultaneously reducing associated costs and simplifying the work of data engineers.

Services like AWS Glue enable the setup of automated data pipelines with advanced data-cleansing and data-transformation capabilities, even in real time. Through technologies such as Amazon S3/Glacier, the Cloud facilitates the creation of durable, scalable data lakes and data warehouses, which allow large volumes of raw data from multiple sources to be centralized at low cost, ensuring easy access and management. Data engineers can easily configure ingestion pipelines that capture data in real-time or batches, ensuring continuous availability accessible to all teams that need it.

In an on-premises environment, ensuring consistent and secure access to data would require considerable infrastructure effort. In the case of process, design, or conceptual errors, the cost of failure would increase significantly. Thanks to the Cloud's pay-as-you-go model, the risk of over-provisioning resources is eliminated; this is a common issue in on-premises systems where failure leads to a significant waste of pre-purchased infrastructure resources. This efficiency enables rapid testing of ingestion processes in temporary environments, meaning pipelines can be iterated easily, errors can be corrected without permanent impacts, and costs associated with misconfigurations are contained.

One of the Key Pillars of Data-Driven Projects: Data Governance

Data governance is a cornerstone of data-driven projects, enclosing various aspects of security, access, and regulatory compliance. It involves managing and controlling the entire data lifecycle, including access authorization, traceability, and activity monitoring. Cloud computing plays a central role in optimizing data governance, helping to manage costs associated with potential failures in the Data Management phase.

In an on-premises context, data governance is often complex and costly to manage, requiring dedicated infrastructure, specialized security teams, and manual configurations that increase both time and expenses. In contrast, the Cloud provides companies with access to integrated governance tools that simplify policy definition, permission

management, and activity monitoring, significantly reducing the risk of errors. Platforms like AWS offer services such as Identity and Access Management (IAM), advanced encryption (KMS), and access logging (CloudTrail), which centralize and simplify permission management.

As already mentioned, one of the primary strengths of cloud environments is process automation, especially for governance. For example, with audit logging services like AWS CloudTrail, sensitive data is continuously monitored, and every access is recorded to address any anomalies or policy violations promptly. This automated traceability facilitates compliance and minimizes the financial impact of errors or misconfigurations. With the Cloud, access issues can also be resolved quickly without compromising data security or requiring manual intervention.

The Cloud in Support of Strategic, Data-Driven Decision-Making

The concept of *Data Value* refers to the process by which raw data is transformed into meaningful, actionable insights that are essential for guiding strategic decisions. However, incorrect analytical choices or ineffective designs can lead to high costs, extend development timelines, and, in the worst cases, result in the failure of the entire project.

In the design phase of an analytical workflow, the data scientist or analyst must make critical decisions on data handling: choosing which machine learning models to use, which variables to include, and how to manage model training and validation processes. With the Cloud, it's possible to test multiple analytical model configurations in parallel, leveraging scalable processing and paying only for usage time. Tools like Amazon SageMaker, for instance, enable rapid, iterative experimentation, lowering the cost of errors or incorrect hypotheses and continuously optimizing models.

The Cloud also facilitates data flow management through advanced orchestration tools, such as AWS Step Functions, which allow for the design of modular and flexible data processing pipelines. In an on-premises setting, errors in the flow would often require redesigning the entire process, leading to time and resource losses.

This modularity allows data flows to be quickly adapted to business needs, providing flexibility even in the case of strategic changes.

Conclusion

The Cloud enables projects that would otherwise require overly complex or expensive infrastructures if developed on-premises. By providing on-demand and pay-as-you-go access to resources, the Cloud paves the way for initiatives such as low-cost disaster

recovery, massive storage, artificial intelligence, genomics, and environmental simulation, all requiring intensive and/or scalable processing.

In data-driven projects, in particular, the Cloud simplifies data management with integrated tools that support data ingestion, governance, and data quality.

The agile, experiment-driven approach centered on the *fail-fast* concept reduces economic and complexity barriers, making ambitious projects feasible for businesses of any size.

About Proud2beCloud

Proud2beCloud is a blog by beSharp, an Italian APN Premier Consulting Partner expert in designing, implementing, and managing complex Cloud infrastructures and advanced services on AWS. Before being writers, we are Cloud Experts working daily with AWS services since 2007. We are hungry readers, innovative builders, and gem-seekers. On Proud2beCloud, we regularly share our best AWS pro tips, configuration insights, in-depth news, tips&tricks, how-tos, and many other resources. Take part in the discussion!

Nicola Ferrari

Cloud Infrastructure Line Manager @ beSharp and AWS authorized instructor champion.I live my life one level at a time getting superpowers by collecting caffeine hidden here and there in my daily map. I'm a hardened internet surfer (yes, I surfed the whole internet... twice!) and tech-addicted with a passion for computers and networking. Building great IT things all nice and tidy contribute to achieving my main goal: the pursuit of perfection!
