Big Data User Forum #2: User Survey Report

10 December 2024

Complete results: https://indico.cern.ch/event/1468866/manage/surveys/5910/results

---

Introduction

The survey aimed to gather insights into current usage, satisfaction levels, user challenges, and expectations for the evolution of the Hadoop service. The responses provide valuable information to enhance the service and align it better with user needs.

---

1. Current Usage and User Profiles

Usage Patterns

- **Frequency of Use**:
    - 50% of respondents access the cluster **daily**, highlighting its integral role.
    - Others use it less frequently, such as **monthly (21%)** or **weekly (7%)**.
- **Clusters in Use**
    - **ANALYTIX** is used by 64% of respondents.
    - **NXCALS** serves the rest, this suggests varying requirements based on cluster.

Components and Tools

- Users rely heavily on a mix of components, with **Spark and HDFS** being the most common.
    - Some specialized configurations include YARN, HBase, and standalone HDFS.
    - **Spark** is the most widely used processing tool, central to most workloads.

Critical Services and Downtime Tolerance

- **Critical Use Cases**:
    - 71% of users' applications are not tied to critical services.
    - However, for the 29% with critical workloads, high reliability is paramount.
- **Downtime Tolerance**:
    - 50% allow extended downtime, while others require stricter limits (under 1 hour).

---

2. User Satisfaction and Feedback

Satisfaction Levels

- **93%** of respondents are either "Very satisfied" or "Satisfied" with the cluster's performance and availability.

- Positive feedback emphasizes **stability improvements** and **robust documentation**.

What Users Appreciate

- **Cluster stability and performance**.
- **Ease of use** and **transparent interventions**.
- Users highlighted "excellent service" and "responsive support" as key strengths.

Challenges Faced

- **Initial Onboarding**: Many users find it challenging to get started.
- **File Duplication and Storage**: Issues with file redundancy lead to inefficient storage use.
- **Access Control**: Limitations in access control hinder use cases.
- **Monitoring Tools**: SWAN's limited monitoring capabilities are a pain point.

---

3. Service Evolution Priorities

User Expectations for Evolution

1. **Modernized Infrastructure**:
    - Strong interest in adopting **Kubernetes** for cluster management.
    - Enhanced **object storage solutions** (e.g., S3-compatible interfaces).
2. **Improved Tooling**:
    - Support for advanced SQL engines like **Trino**.
    - Real-time analytics capabilities.
3. **Resource Allocation**:
    - Suggestions to improve **fair-share scheduling** and throughput management.

Other Suggestions

- Address the **disconnect** between service provider perspectives and user needs.
- Provide **targeted documentation** for PySpark and other tools.

Key Feedback on Evolution

- Users appreciate ongoing stability improvements but expect a shift toward flexibility and modern tools.
- Requests for better resource allocation and monitoring highlight the need for greater user-centric enhancements.

---

4. Recommendations

1. **Enhance Stability and Reliability**:
    - Continue prioritizing cluster stability, as it is the most valued feature.
    - Focus on addressing downtime requirements for critical services.

2. **Simplify Onboarding**:
    - Develop beginner-friendly guides tailored to specific use cases.
    - Provide hands-on tutorials for new users.
3. **Modernize Infrastructure**:
    - Pilot Kubernetes support and expand S3-compatible object storage.
    - Evaluate and integrate SQL engines like Trino to enhance data querying.
4. **Improve Monitoring and Tool Access**:
    - Invest in better monitoring interfaces for tools like SWAN.
    - Streamline APIs to facilitate easier integration with external systems.
5. **Strengthen Communication**:
    - Bridge the gap between providers and users by gathering feedback more frequently.
    - Align future developments with evolving user needs.

Conclusion

The survey results confirm that the Hadoop service is highly valued for its stability and performance. However, users also see opportunities for improvement, particularly in onboarding, resource allocation, and tool modernization. By addressing these areas, the service can continue to meet diverse needs while evolving to support modern workloads and tools effectively.