Analyzing Device Usage & Failure Rates for a F500 Data Storage Vendor

CHALLENGES

- Completing projects using existing big data solutions required armies of programmers and consultants, thus reducing ROI and scope
- In-house SQL and Python experts were finding it difficult to move projects forward with other big data platforms
- Information extraction and insight discovery was complex and time-consuming, and resulted in gaps when handling unstructured data

SOLUTION

- With Xcalar's end-to-end platform for the entire pipeline—sourcing, cleansing, and transforming data—for delivery of ad-hoc analytics to business analysts, the analytics team is now able to execute more projects faster
- Xcalar Design empowers DBAs, programmers and analysts to do their data prep and analytics using familiar skill sets of SQL, Python, and visual programming
- Xcalar pulls in raw data from diverse sources—unstructured, semi-structured and structured—and allows users to perform complex relational operations on it

BENEFITS WITH XCALAR

- Reduction in time-to-value **from 3 months to 4 days**
- Increase in developer productivity by **10x**
  - Enabling users to expose, assess, and resolve data anomalies at every step achieves higher accuracy
  - Full data lineage helps improve data quality and governance
  - Modular dataflows and UDFs make code reusability and maintainability easier
- Improvement in query performance by **200x**

Introduction

A leading computer hardware company collects telemetry data from most of their systems deployed around the world to better understand how their products are being used and how reliable they are. Data bundles arrive daily from several hundred thousand systems, resulting in petabytes of structured, semi-structured, and unstructured raw data that must be stored, managed, and analyzed. Xcalar partnered with this company's data analytics team to improve the efficiency and completeness of their data workflow.

Challenges

The size and complexity of data bundles required a complex toolchain to process and analyze the data. A typical bundle consists of more than 100 sections, each represented by a single file. These files come in a wide variety of formats, including XML, Excel, binary, logs, text-based tables, and free-form text, with some containing more than one character set. File sizes range from a few kilobytes to over a gigabyte. Field counts per file type vary from a handful to over 500. Formats vary widely across system types, product models, and software versions. It was a challenge to gain insights from this extremely diverse data, determine the root causes of system problems, and predict future problems for the following reasons:

- **Required Tool Expertise:** Specialized knowledge of each product in the toolchain was required which involved developers in different groups across various geographies.
- **Complex, Frequently Modified Data Bundle Formats:** Complex parsers with thousands of lines of code pre-processed the raw data bundles into intermediate on-disk tables. Few people understood these transformations and how to maintain them to keep up with frequent format changes and enhancements.
- **No Process Control:** Neither were there any mechanisms to ensure process accuracy, completeness and record uniqueness nor were there any views of lineage back to the original data.
- **Little Code Reusability:** Analyzing usage trends and predicting failures based on the complex source files required extensive programming with various tools. Code reusability, iteration, and debug cycles were lengthy and inefficient without a modular framework.

Xcalar Solution

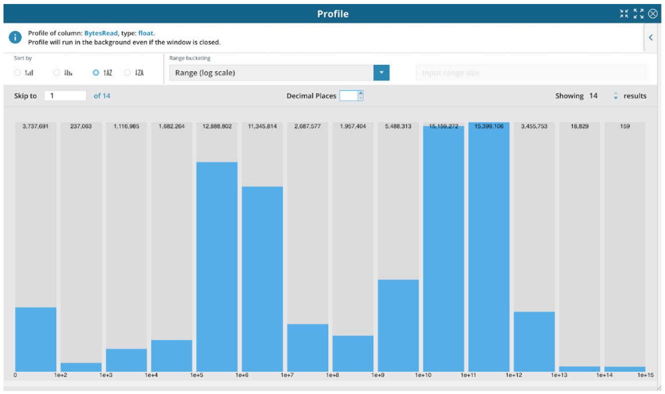
The solution leverages many features provided by the Xcalar Design visual interface, including shared datasets and UDFs (user-defined functions), custom Python parsers, profiling and statistical analysis, data lineage, and auditability. Xcalar Data Platform provides the scale and performance to model dataflows on multiple datasets with hundreds of millions of rows. Users can then operationalize these dataflows for the entire petabyte-size datasets with a few clicks.

Parsing Data Bundles

Users begin modeling by importing files containing data bundle sections from NFS servers to create datasets. Then, as they need further information, they create additional datasets and create joins based on the unique data bundle identifiers. Xcalar natively imports bundle sections in open formats, such as XML, Excel, CSV, JSON, Parquet, or text. Users quickly develop short import UDFs in Python to parse custom file formats, such as tables-as-text or key-value pairs interspersed with unrelated text. In this fashion, the unstructured data is either imported as-is or is first refined into semi-structured or structured data and then imported. Because parsing is performed on one data bundle section at a time, parser code is modular, simple, and reusable. Because of Xcalar's True Data In Place™ technology, the original data files remain unchanged throughout; they are simply referenced as needed and no intermediate tables are required to be written to disk. As bundle section formats evolve over time, parser code is modified or updated directly in Xcalar Design.

SQL, Visual programming and Python for Data Profiling, Exploration, and Validation

After users create datasets from the underlying raw data, they use Xcalar Design to understand and validate the data and build their dataflows using SQL, visual programming and structured programming. Xcalar's powerful scale-out compute engine enables Xcalar Design users to play with tables of hundreds of millions of rows, and quickly validate, explore, and transform weeks' or months' worth of bundle data interactively. Because data lineage to the original raw files is preserved, it is easy to determine the root cause of data errors and resolve them.



Profile

Profile of column: BytesRead, type: float.
Profile will run in the background even if the window is closed.

Sort by: ○ ○ ● ↑AZ ○ ↓ZA

Range bucketing: Range (log scale)

Input range size

Skip to 1 of 14

Decimal Places

Showing 14 results

Figure 1 shows an example Profile Graph. The graph shows the distribution of the number of bytes read from a large number of disks, using a log scale.

A comprehensive set of visual tools such as the Profile Graph shown in Figure 1, in combination with SQL and Python, allows users to immersively pursue insights from the data at a level of productivity that is dramatically better than what was possible with the company's incumbent analytics toolset.

Machine Learning Algorithms

Developers use the embedded Jupyter Notebook to integrate and train TensorFlow ML algorithms against billion-row data sets in real-time. Xcalar dataflows for the predictive analysis of device failures execute regularly to enable the optimization of spare parts inventories.

Operationalizing Dataflows

Once dataflows are successfully modeled, they are optimized for production and scheduled to execute regularly. A prime example is a monthly dataflow that analyses time-series device data joined with customer contact data to highlight the customer-initiated downgrades in the past month. This insight enables them to contact their customers to understand the reasons for the downgrades and respond accordingly to improve product quality and customer relationships.

Key Features

- Point-and-analyze capability for any source datafile–unstructured, structured or semi-structured
- Ability to import data from custom sources in multiple parts (sections), each as a separate dataset
- SQL, visual programming, and Python for code development flexibility
- Integrated machine learning tools, including Tensorflow

About Xcalar
Xcalar is a cloud-scale platform for data processing applications. It provides an IDE for developing sophisticated business logic, and a flexible computing framework—for querying and relational compute—directly on the data lake. Companies of all sizes use Xcalar's versatile IDE that supports open standards and extensibility to interactively build dataflows using SQL, visual programming, and structured programming, then execute them at petabyte scale. Xcalar's enterprise-grade software scales linearly across hundreds of nodes and thousands of users for both public/private cloud and hybrid deployments.

Xcalar, Inc. 3031 Tisch Way Suite 450 San Jose CA 95128 USA Tel 408-471-1711 www.xcalar.com

Xcalar shall not be liable for any technical errors in this document. This document is subject to change without notice.
XSB-96989-14-14-02