SpringDB

47 Hurdles on Your Roadmap to Working with Data Cubes

“Ginormous Datasets Are A Massive Pain In The Butt Even For The Most Experienced Teams”

Initial Assessment

1. Initial Assessment: Knowing where to start with the data
2. Data Volume Uncertainty + Estimation: Having no prior knowledge of the number of files or their sizes, making it difficult to prepare for data receipt, ongoing storage, and processing.
3. Build timeline to show all processes from initially creating/linking cloud buckets to importing data in a database.

Receiving Data

4. Cloud Bucket Authorization: Understanding the authorization process to access the bucket is challenging. Setting up shared keys with the data vendor (typically 3-4 day turnaround)
5. Initial Zip File Size: Encountering an unexpectedly large initial zip file of multi-terabytes (depending on the subscription) leads to handling and storage challenges.
6. Extracting Zip Files: Discovering that it is CPU and storage-intensive to extract terabytes of zip files, complicating the data preparation process.
7. Unzipped CSV File Size: Dealing with the unzipped CSV files, which expanded to a factor of 1.5x the original size, adding to the storage and processing complexities.
8. Watchdogs - Creating and setting up a watchdog process to monitor new zip files as they drop to cloud buckets. Similarly, creating and setting up a watchdog process to unzip CSV files.

Data Volume + Scope

9. Understanding Data Scope: Grasping the vastness and specificity of the data provided.
10. Data Volume: Handling the sheer volume of data efficiently.
11. Understanding Data Schema: Facing difficulties in comprehending the data schema and object relationships provided by the data vendor, which is crucial for effective data utilization.
12. Create schema for CSV files
13. Deciphering Complex Data Joins: The complexity of managing and understanding how to join various data tables provided by the data vendor for comprehensive analysis.
14. Reorganizing data: More effective data organization, integration, and analysis, directly impacting the utility of the data.

Data Storage

15. Cloud vs. On-Premise Data Storage: Deciding between cloud and on-premises, considering costs and scalability, and balancing cost, efficiency, performance, flexibility and scalability.
16. Database Storage Requirements: Uncertainty about the scale of storage required for
17. efficiently managing the data with the cloud storage provider, including decisions on storage type. Provision storage and download monthly and quarterly tables.
18. Cost Management: Balancing the costs associated with storage, network, and compute on AWS, Snowflake or other cloud vendors, especially when considering the volume of data.
19. Deciding on Storage Type: Make informed decisions on the type of storage needed with your cloud vendor to accommodate the data efficiently and cost-effectively.
20. Learning Curve with S3/Snowflake: Using S3, Azure, GCP, Databricks, Oracle, or Snowflake for data storage and management is a challenge if you don’t have initial familiarity with these solutions.

21. Cloud Provider Costs: Snowflake, AWS, GCP, Databricks, and Oracle are versatile but extremely expensive at scale, which can be a massive surprise and require budget adjustments.

Data Cleaning + Preparation

22. Loading CSV Files: Challenges with loading CSV files efficiently into databases due to size and format complexities.
23. Optimizing Loader for CSV Ingestion: Figuring out an effective way to ingest CSV files, requires understanding and optimizing of the data loader process, especially for those tables that exceed 1+ billion records daily.
24. Data Quality Issues: Addressing issues with jagged edges in the data caused by unexpected characters and formats that impact data loading.
25. Unexpected Character Issues in Data: Dealing with unforeseen characters within the dataset that caused loading errors and data integrity issues.
26. Data Cleaning and Preparation: Extensive data cleaning may be required to ensure data quality and usability for analysis purposes (depending on needs). Data blending, standardization, cleanup, data transformation needs to be performed.
27. Data Transformation for Analysis: Transforming vendor data into formats suitable for analysis and reporting presented technical challenges. Extensive work denormalizing tables, partitioning tables and adding indices for performance must be completedData Duplication and Redundancy: Identifying and preventing unnecessary duplication of data, which could lead to increased storage costs and complexity.

Working with Data

28. Efficiency in Query Execution: Struggling to write and execute efficient queries that could handle large volumes of data without incurring high costs or long processing times.
28. Opportunity to minimize costs and maximize insights drawn from the data. Can your designed workflow download, unzip, load, and process your large datasets in less time than it takes for new data to be published?
29. Complex Data Joins: Handling complex data joins for comprehensive analysis.
30. Automating Data Processes: Finding efficient ways to automate repetitive data tasks to save time and ensure consistency in data handling.
31. Cloud Services Cost Management: Addressing the financial implications of utilizing cloud services for data storage and processing, aiming to optimize and reduce expenses.

Data Across Time

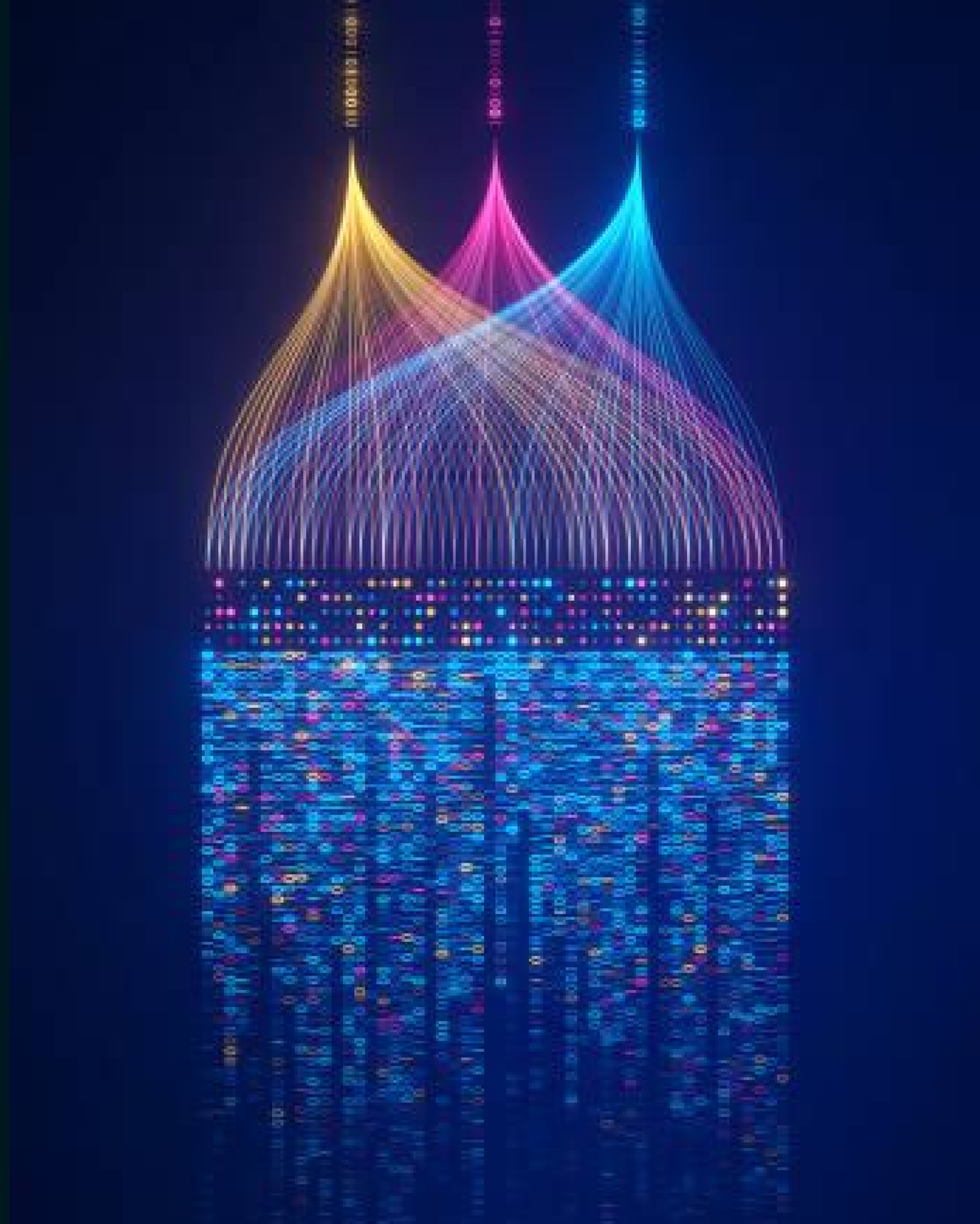
32. Historical Data Analysis: Effectively using past data.
33. Determining Data Refresh Rates: The uncertainty surrounding how frequently the data would be updated and the implications for storage and processing.
34. Retention: Determining how long do you need to keep your data, ensuring you have the storage and processing requirements
35. Real-time Data Processing Needs: The need to establish a system capable of handling real-time data processing and analytics is a huge challenge.
36. Real-time Analytics Capabilities: Establishing systems and processes capable of supporting real-time analytics for timely insights, which was particularly challenging given the data volume.
37. Not all data is equal: Establishing different methodologies for both real-time processing and historical data analysis and catering to different analytical needs can maximize the data's utility.
38. Scaling – provision compute and storage instances for scale
39. Backup and recovery – Perform and manage schedule for incremental and full database backup and recovery procedures

Data Integration + Enrichment

40. Legacy Systems Compatibility: Had to ensure that the new data and tools were compatible with existing legacy systems to avoid operational disruptions.
41. Managing Data Across Multiple Platforms: Ensuring seamless management and accessibility of vendor data across different platforms and tools used by the team.
42. Identify Custom Data Requests: Facilitating specific data requests that may not be readily available.
43. External Data Synchronization Challenges: Synchronizing external data sources with vendor data enriches and enhances the dataset for more comprehensive insights.
44. Data Enrichment: Augmenting vendor data with additional first and third-party data sources.

Scalability + Long-term Planning

45. Scalability of Data Solutions: Concerns over whether the chosen data storage and processing solutions could scale effectively with the growing volume of data.
46. Strategic Long-term Data Planning: Developing a strategy that could adapt to and accommodate evolving data needs, including storage, processing, and analysis.
47. Data Lifecycle Management Considerations: Effectively managing the stages of data from creation to deletion, which involves considerations for storage, accessibility, and compliance.



OVERCOMING CHALLENGES WHEN RESOLVING IDENTITY

Data Needs Assessment Consultation

PLAN
Assess, Decide and Prepare

RECEIVE
Obtain, Process and Store

USE
Analyze, Optimize and Update

EXPAND
Connect, Enhance and Augment

When you engage with our DNA Consultation we will help you identify specific nuances of your situation and create a bespoke data onboarding and management plan for working with 3rd Party Data.

At the end of the DNA Consultation, you will walk away with a completely custom Data Implementation Blueprint “DIB” that will be your roadmap for getting things done time-, cost- and resource-effectively.

SpringDB

Data Needs Assessment Consultation

IB will give you and your team detailed guidelines, actionable steps, and recommendations based on YOUR assessment results and YOUR situation (Note: this is not a ‘cookie cutter’ template). Here's a complete breakdown of what the DIB includes:

1. **PLAN**
   This phase involves evaluating the size of 3rd party datasets at a high level and how it would work with your existing systems. It includes getting familiar with the data structure, review of hardware and software requirements, and understanding the specific nuances of working with really huge datasets.

   Key activities typically include identifying stakeholders and human resources, planning implementation, conducting readiness and infrastructure analysis, and outlining a roadmap for using the data.

2. **RECEIVE**
   In this phase, the focus is on getting vendor data into your organization's database or data warehouse. It involves setting up data pipelines, ensuring data integrity during various transfers, and technical preparations for storage and processing.

   Key tasks typically include access to data, downloading and unpacking data, data mapping and schema design. At this stage a loading process is created and tested resulting in the data being placed into your database to make the data ready for analysis and next steps.

3. **USE**
   This phase is about maximizing the efficiency and performance of data processing and analysis. It includes fine-tuning data pipelines, optimizing query performance, and enhancing data structures for faster and better access and reduced costs.

   Key activities typically involve creating custom indexes, partitioning data for better management, implementing caching mechanisms to improve data retrieval speed and more.

4. **EXPAND**
   In this final phase, the focus shifts to enriching vendor data with additional internal & external sources or supplementary datasets. Either by leveraging crawling and scraping capabilities (offered by SpringDB) or by integrating data sources already available to the customer (BYOD) as well as new ones that can be sourced out through the SpringDB data marketplace.

   Key tasks typically include developing web scraping scripts to extract relevant data from online sources, integrating crawled and/or in-house data with vendor datasets, sourcing additional data sets, and ensuring data quality and performance during this expansion process.

If time to market is critical, we guarantee that DIB will significantly streamline your onboarding process from 3+ months down to 1-2 weeks. The DIB not only helps you and your team overcome the initial hurdles of working with vendor datasets but also sets you up for long-term success.

Complimentary Discovery Call

If working with SpringDB sounds interesting, let's schedule a Discovery Call to see if we are a fit. The call is free, and even if we are not a fit, you will walk away with greater clarity about your next steps working with any vendor’s Data.

Data Management Solutions from SpringDB

We help big data buyers navigate all of the technical complexities and hurdles of acquiring, integrating, and deriving value from 3rd party data.

1. Data Needs Assessment “DNA” Consultation

This phase involves evaluating the size of 5x5 data at high-level and how it would work with existing systems. It includes getting familiar with the data structure, review of hardware and software requirements, and understanding the specific nuances of working with really huge datasets. Key activities typically include identifying stakeholders and human resources, planning implementation, conducting readiness and infrastructure analysis, and outlining a roadmap for using the data.

3. Cloud Access to Data

Comprehensive and versatile access to your processed 5x5 data tables via cloud with one or more of the options below::

- Cloud CSV – processed and transformed 5x5 CSV files hosted on SpringDB or AWS S3, ready for easy integration into your existing solution.
- Cloud SQL – MYSQL instance with processed and transformed 5x5 data tables, ready for use immediately
- InstaDBTM – full access to cloud-based processed and transformed 5x5 data tables via Web UI and/or API at minimal cost and stellar performance (compared to Snowflake and similar options)

2. Preparing and Managing Data

so you don’t have to. We provide ongoing data processing and management services that produce “clean” cloud-hosted tables that you can access immediately from almost any environment (in contrast to raw ginormous zipped-CSVs you can’t work with directly)

4. Pro Services

ad-hoc and custom data requirements, including:

- Data enrichment and linking to additional data elements and data sources
- Integrating third-party platforms and storage, cloud, data and AI solutions
- Advanced and volume data crawling and customized scraping
- Complex queries and data pulls and extracts
- Scaling up (or down) in infrastructure
- Data-related cost reduction and optimization services
- First line of support for data concerns and more…

SpringDB