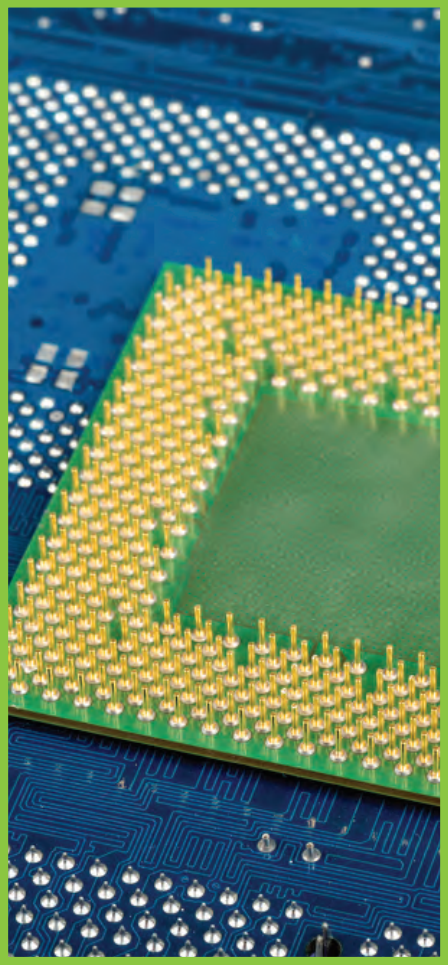
Case Study

Technology Company Masters Big Data with Cloudera & Master Data Maestro



Powering Value Cost Chain at Heart of Analytics Hub

Business Challenge

Over decades of developing industry-leading technology components, this Company has demonstrated an unwavering commitment to delivering technology and manufacturing leadership to Original Equipment Manufacturers (OEMs) to power a wide variety of devices. This presents a data management and analytics challenge: How could the Company gain intelligence on how their *components* were performing based on sales of the OEM *devices* powered by those components?

This was the primary driver behind the Company's integrated analytics project, which was designed to provide information for the Company's business analytics teams on how their components are performing based on device sales data. While they knew how many units of a specific component they had sold to a given manufacturer, they needed to know how many of the finished products that carry the Company's components were in turn shipped by that manufacturer, and how many of those were then sold in the marketplace. They also sought to understand how the manufacturers were positioning the finished devices in order to identify emerging markets, evaluate how their company is doing, track what their competitors are doing, and ultimately determine how to focus their time and effort to secure and maintain competitive advantage.

In order to provide this kind of actionable intelligence to their business analysts, the Company relies on reports from a variety of external sources, each of which provides information on device consumption, from the retailer perspective. These reports are

MDM Domain
Customer
Product

Industry
Technology

Profisee®

presented in various formats, with different structures, nomenclature, field attributes, and so on. Each source – and the Company uses 30 to 40 different sources – provides one part of the sales and marketing performance picture. The challenge was to consolidate the data from across these disparate sources to get the whole picture, and allow their analysts to view it by specific device, country, currency, and device type, to develop strategic intelligence.

The sales information provided by these third-party sources for various retailers contains information on devices containing the Company's components, as well as those of their competitors; basically, all of the information at point-of-sale regarding what is being sold by a given retailer. This not only enables the Company to look at how their products are performing in the market, based on the devices being sold, it also has the potential to give them a deep insight into their competitors' sales. With the right data management capabilities, these reports can be leveraged to identify a competitor's push into a new market, or understand why their sales of a specific device configuration exceed the Company's. This is key intelligence to fuel strategic product development and marketing initiatives – but it requires that the analytics solution be able to manage and consume the full range of information embedded in the third-party-provided sales data. And that means the Company must be able to accurately track the data by *device*.

Further complexity is added by the fact that the sales data coming in is about the same devices, but they are inconsistently referenced, and can't automatically be reconciled to the same device identifier. They needed to be able to incorporate the data from these sources into an OEM product master.

Evolving their consumption measurement capabilities and improving utilization of retail sales datasets were high strategic priorities for the Company's analytics project. Specifically, the project sought to implement an architecture built around Cloudera's analytic data management capabilities , but they needed a solution for mastering, mapping, de-duplicating, and cleansing the data, and ultimately conforming it to the Company's standards. Further, they needed the solution to be seamless, and able to evolve at the pace of the Company's sales and marketing efforts to support potential transformation and growth. That's where Profisee's Maestro Master Data solution comes into play.

Accommodating New Data Sources on Demand

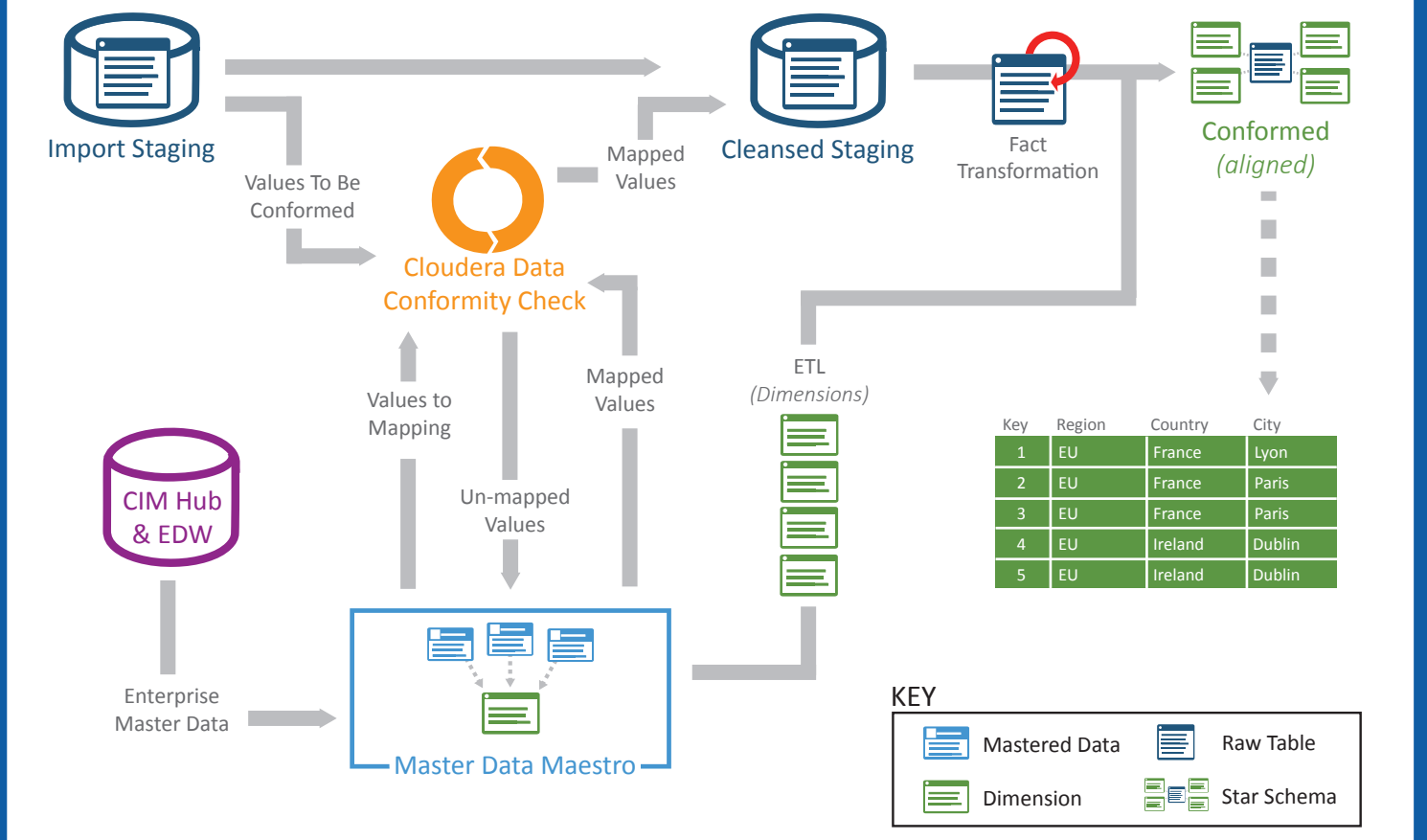
A solution was also built into the initial design of the Maestro implementation to accommodate processing of an information set that the company has not seen before, from a schema perspective. For example, they receive an Excel file from a new source, and they want to conform the data it contains into their standards. Cloudera will interrogate the new file for its schema, and identify for Maestro the different pieces of data that they want to conform against; for example, "conform information contained in Column F against country." Maestro will then spin up, on demand, an entity based upon that structure, with all of the attributes the company wants to conform against, with the conforming status of each attribute. Maestro then processes the data as it would for other, known, sources.

Maestro Solution

Building the analytics solution on top of Cloudera gave the Company the big data computing power needed to incorporate the high volume of retail sales data into a value cost chain. Integrating Maestro into the solution added the needed master data, stewardship and governance support.

When they started the analytics project, which was originally based on Microsoft SQL Server Parallel Data Warehouse (PDW), the Company had been talking with Profisee about a hierarchy management solution. When the decision was made to move from PDW to Cloudera, Profisee rose to the challenge, demonstrating that Maestro could integrate readily with Cloudera, and

Analytics Hub Architecture



provide the master data functionality they needed.

Under the analytics hub data architecture, Cloudera receives "raw" data from a variety of external sources through a custom import process. The raw tables are staged in Cloudera, and the values to be conformed are run through a Cloudera data conformity check, which boils down the data, doing the basic matching and mapping, and getting rid of obvious duplicates (exact matches) in the data. Cloudera then passes to Maestro any unmapped values, data without any obvious matches. Maestro takes it the next step, running sophisticated "fuzzy" matching algorithms and matching strategies on the data to master it at a level of detail far beyond what Cloudera is able to do.

Cloudera also sends Maestro the master schemas for particular attributes they want to align or conform against (e.g., country, currency, device type). For entities where there is a defined master list, Maestro relies on that list as the source. However, the Company had no internal master list of devices (OEM products), so Maestro generates those masters using synonym lists and attributes, serving as a kind of 'Rosetta Stone' for deciphering the device data, regardless of source, aligning it all into a common language.

For example, one list (or report) from a single source may contain information for every store in the country for the "ABC Inc." large format retail chain, reporting on each store's monthly sales volume for a specific device containing one of the Company's components – we will call it "Mobile Device A-1".

Initially, this information comes in to Cloudera, which processes the data for exact matches: "Have I seen this device before?" If it is identified as Mobile Device A-1, as an example, Cloudera says, "I've seen this device before, I've seen it from this source before, and I can match it with the corresponding master record, as the Company has defined it for this device."

In this way, Cloudera performs an "exact match" against identified masters, but it lacks the ability to perform

fuzzy matching, to master the data, or to adjust or optimize matching strategies to get data in the needed format. So on preliminary check-in of the source data, if no exact match is found, Cloudera passes the data to Maestro, which does a compare for value match.

The challenge is that the model name, as listed by different data sources, is unpredictable and inconsistent – Mobile Device A-1; MD A-1; MobileA1; etc. So to match it, Maestro looks at the attributes of the device: What type/size of device is it? What component is running in it? Who is the device manufacturer? If the data can be matched and mapped in this way to an existing value, the mapped value is returned to Cloudera to be loaded into cleansed staging for schema embedding.

If Maestro identifies new master values from the data, these are also passed back to Cloudera to be mapped, and become part of the set of master entities Cloudera has available for matching from that point forward. This set of mastered values in Cloudera grows through the matching activities of Maestro, so if new data comes in that exactly matches the expanded set of master values, it does not need to be passed to Maestro again.

This allows the different elements of the analytics solution to focus on their respective strengths, continuing to improve the speed and accuracy of the system over time. Cloudera and Hadoop can handle processing the big data on the front end, where the Company is receiving around 20 million records at once, from each single source, with around 30 different sources total. Cloudera is well-suited to perform the initial data staging and conformity checks at this level of volume, and only has to pass to Maestro the subset of, for example, 500,000 records that are new and unique, that have never before been seen or mapped.

Maestro can easily process that volume, and provide results back to Cloudera in a very timely fashion. And, as time goes on, the processing speed increases, because every new value that is mapped by Maestro and passed back to Cloudera builds up the knowledge base of match groups. As a result, the Company is developing a master matched-pair list, and they'll be able to compare it to data coming in from any source, even if it is in a slightly different form. This will enable

Maestro Matching & Mapping Example

The first time a report comes in from large format retail chain, containing country name "Brasil," spelled with an "s," Cloudera attempts to conform the country information to its defined master data, which contains country name "Brazil," spelled with a "z." Unable to find an exact match for "Brasil", Cloudera packages it up, along with any other values for which there are no exact matches, and sends them on to Maestro.

Maestro has a master record which contains the country name "Brazil" and the country code "BZ." The new value to be matched is loaded from the source file, with a specific source identifier. Now there is a master record with a country name value of "Brazil," and a source record with a country name value of "Brasil." Maestro matching runs against the master, groups the source record with the master of Brazil spelled correctly, and then harmonizes the name field into the new record that just came in. Now there is a matched pair, with Brasil misspelled and Brazil spelled correctly on the exact same member all the way across. Maestro then writes that information back to the results table that spun up for the new file coming in.

Cloudera polls those results daily. In this case, it receives the mapped data – "Maestro received 'Brasil' misspelled; here's the value properly spelled; here's the master code for that particular value" – and incorporates the matched and mastered information into the mapped values in Cloudera. In that way, when a new file comes in with the country name of "Brasil," it doesn't get passed to Maestro again, because Cloudera filters it on the front-end, based on previously mapped match results, and only passes to Maestro the changes and items Cloudera has never seen before, for which there is no exact match.

the Company to start seeing broader patterns in the data as it comes in; if the system data sources are providing the same type of device information, they will be able to recognize that, and utilize that perspective to better mine or apply the data for BI analytics.

Creating dimensions for BI

For all of the entities Maestro has created, it has created hierarchies, as well. For example, 'country' rolls into a 'sub-geo,' that in turn rolls into a 'geo.' Cloudera takes all of the master entities Maestro creates, for example, 'country,' and creates dimensions from those. In addition, Cloudera flattens any of the hierarchies Maestro creates into a dimension table, and uses these to build tabular models, making those dimensions readily available for BI analysis.

Hierarchy Management

In addition to externally-sourced data, Maestro works with two different sets of "master" data: *Enterprise masters* and *business segment masters*. Using an example of a country listing, we can see how the hierarchy management requirements differ between enterprise and business segment masters.

There are more than 250 named countries on the enterprise master list that everyone in the enterprise agrees upon. However, individual business segments may need to identify, sort, or group the countries in different ways for their work; for example, rolling Jamaica sales volume into a larger "country," identified as Caribbean Islands.

To accommodate this approach, the Maestro team created a *Master Data Type* attribute that defines whether the entity is enterprise or business segment master data, and, for the latter, another attribute that identifies which business segment it belongs to. This allows the Company to create functional designations such as "Caribbean Islands" inside the country entity.

While Intel's business rules require hierarchies to have a four-level structure – Country entities roll up to a Sub-Geo, to a Sub-Geo Grouping, and to a Geo – business segments are allowed to skip levels, and each hierarchy in each business segment can roll up into a different structure. This means that each of the business segments can build and manage the hierarchies in the manner that works best for them.

This capability is supported by the creation of attributes inside of each entity for a group's specific hierarchy. For example, in *Country*, there is an attribute called *Sales & Marketing Group Geo Hierarchy*, which contains the pointer to the *Sales & Marketing Group Sub-Geo*. And in that Sub-geo, there's an attribute that is the *Sales & Marketing Country* hierarchy.

By creating attributes at each level of the Geo structure they want to use, each group can add their own Country roll-up hierarchy. These attributes all point to a *Geo Hierarchy Type* entity that tags it for that specific group. So, from a view perspective, if you roll up through just the Enterprise attributes, you get an Enterprise hierarchy view; if you want a Sales & Marketing view, you roll up through the Sales & Marketing attributes, and see that hierarchy.

This approach honors the Company's preference not to use explicit hierarchies, and still gives them the flexibility needed to meet the needs of different working groups – flexibility and control that they didn't have before the Maestro implementation. It makes the data more accessible and usable for the various business segments.

Results

Cloudera enables the Company to handle the big data they are accessing through their various external source reports, but without incorporating Maestro into their analytics implementation along with Cloudera, they would not have the innate ability to master and manage their data.

With Maestro in place, they gain the advantage of really understanding what actionable information is available to them from within the big data streams. Prior to the Maestro implementation, the data managers might get the data for a given month, and it could take them two to three weeks to sit down and process that data, and try to start putting pertinent records into the system. And that's only allowing them to look at a few hundred of their top device records.

Maestro can process that same month's data, based on all the different attributes defined for matching and mastering the device data, in about an hour-and-a-half, and do it in such a way that they get back data on some 88,000 records. How much of the information is incorporated into their analytics will be determined by specific strategic business requirements and resource thresholds, but Maestro enables them to see all of the potential business intelligence available to them from their external data sources – to see how big their big data really is, in terms of potential for action.

By the same token, out of a typical record set that comes in with roughly two million records, Maestro is able to reduce that to around 100,000 masters that the data stewards will deal with. They can then slice and dice that data as desired – by OEM or by specific attributes, for example – and break down the data even further. The data stewards do that within Maestro, manipulating the data, getting it into the right format, then providing it back out to the rest of the organization through Cloudera, through data cubes and dimensions, flattened hierarchies, and so on, for consumption in BI initiatives.

Without Maestro, the Company would still be able to deal with large volumes of data, but they wouldn't have that one central source, that key element that allows them to deal with the process flows and the structure surrounding the master data itself. And, it gives them a primary mechanism to improve their data governance, because now they can see and deal with the actual master data coming across, and manage decisions about what gets designated as a master.

With Maestro in place within the analytics solution, the project teams, data managers, and data stewards can utilize the MDM toolsets to give the business users within the Company the data they need – conformed, clean data that's been aligned to the Company's standards, so it can be merged and analyzed from a common point of view across the entire organization.

Profisee — A Trusted Advisor

Profisee is a master data management software company focused on delivering enterprise-grade MDM capabilities through its Master Data Maestro software suite. As a Microsoft Gold Application Development Partner, Profisee has a worldwide reputation for Master Data Management expertise and competence with Microsoft Master Data Services.

website :: www.profisee.com
Americas :: +1 678 202 8990 | info@profisee.com • Asia & Oceania :: +61 (0)2 9931 7874 | aus-info@profisee.com
UK :: +44 (0) 2084 336572 | uk-info@profisee.com • EMEA :: +44 (0) 2084 336572 | emea-info@profisee.com