O'REILLY®

Integrated Analytics

Platforms and Principles for Centralizing Your Data

Courtney Webster

Strata+ Hadoop WORLD

Make Data Work

strataconf.com

Presented by O'Reilly and Cloudera, Strata + Hadoop World is where cutting-edge data science and new business fundamentals intersect—and merge.

- Learn business applications of data technologies
- Develop new skills through trainings and in-depth tutorials
- Connect with an international community of thousands who work with data

Job # 15420

Integrated Analytics

Platforms and Principles for Centralizing Your Data

Courtney Webster

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Integrated Analytics
by Courtney Webster

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com.

Editor: Tim McGovern
Production Editor: Leia Poritz
Interior Designer: David Futato
Cover Designer: Randy Comer

December 2015: First Edition

Revision History for the First Edition
2015-12-15: First Release
2016-02-05: Second Release

978-1-491-95270-2

[LSI]

Table of Contents

Integrated Analytics: Platforms and Principles for Centralizing Your Data

Abstract

Data centralization merges different data streams into a common source through unified variables. This process can provide context to overly-broad metrics and enable cross-platform analytics to guide better business decisions. Investments in analytics tools are now paying back a 13.01:1 return on investment (ROI), with increased returns when these tools integrate with three or more data sources. While the perks of centralization are obvious in theory, the quantity and variety of data available in today's landscape make this difficult to achieve.

This report provides a roadmap for how to connect systems, data stores, and institutions (both technological and human). Learn:

- How data centralization enables better analytics
- How to redefine data as a vehicle for change
- How the right BI tool eliminates the data analyst bottleneck
- How to define single sources of truth for your organization
- How to build a data-driven (not just data-rich) organization

Introduction

Data is a valuable asset and, as a result, companies are more hungry for data than ever before. New products address that need by pro-

viding metrics on every step of a sales pipeline (from social media and marketing, to website traffic, sales and product usage, through customer support). The increase in software as a service (SaaS) products contributes to the data firehose—by 2016, the use of cloud services for business processes will have accelerated past current forecasts by 30%.[1] But more data doesn't necessarily translate into better analytics, given how difficult it is to unify SaaS-based information with other internal and external data streams.

Internal Data Sources	*External Data Sources*
• Engineering/manufacturing • ERP systems • Sales/financials • Support/CRM • Marketing	• Social networks • Clickstreams • Sensors • Websites • Supply chain/logistics

The challenge is no longer to gather more data—it is to build meaningful analytics with various (and highly dynamic) data sources. To tell a complete story, your analytics must utilize more than one data stream. Centralizing your data provides the structure necessary to enable cross-platform analytics.

How Centralizing Data Provides Context for Better Business Decisions

Consider this theoretical example from Colin Zima, formerly of HotelTonight:[2]

1 Daryl C. Plummer, Leslie Fiering, Ken Dulaney, et al., "Top 10 Strategic Predictions for 2015 and Beyond: Digital Business Is Driving 'Big Change.'" *Gartner*, 4 October 2014.

2 Looker Webinar, "5 Ways Centralizing Your Data Will Change Your Business." *Vimeo*, 3 August 2015.

Hotel	*Support Issues*
West Side	200
East Village	50
Financial District	4
Bed & Breakfast	1

You could evaluate the "quality" of a particular hotel based on its total number of support tickets. Based on this data, you may decide that the West Side hotel created the worst experience for your customers. As a result, you decide to pull marketing for this hotel or find an alternate hotel to utilize in this area.

If you were to consider the number of support tickets alongside a different data stream, like the number of bookings, an emergent property of integrated data (which we'll call context) paints a different picture:

Hotel	*Support Issues*	*Bookings*	*Issues per Booking*
West Side	200	100,000	0.2%
East Village	50	10,000	0.5%
Financial District	4	4,000	0.1%
Bed & Breakfast	1	100	1%

You find that the support tickets were actually a small fraction compared to the West Side hotel's total bookings, and now the Bed & Breakfast appears problematic. You'd make a *different business decision* now compared to when you considered each data source independently.

Now imagine you could pivot the data to map support tickets over time:

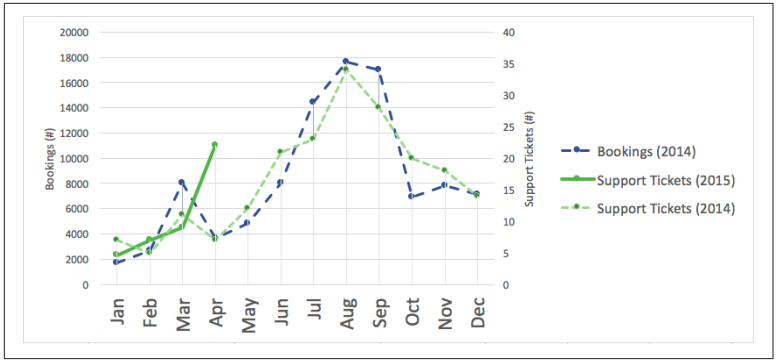

Figure 1. Comparing current support tickets to historical support tickets and bookings for the West Side hotel shows an anomaly

Compared to last year's numbers, you find that support tickets are peaking right now (hypothetically, April 2015) at the West Side hotel and that this peak is out-of-sync with expected seasonal bookings. You drill down into this month's support tickets and find that they point to a rude hotel clerk, whom you promptly fire.

The ability to make this decision relied on a few things:

1. Centralized data, which allowed you to compare two different data streams (support tickets to bookings)
2. Real-time analysis, which allowed you to identify an anomaly before it had a long-term negative impact
3. Drill-down capability, which allowed you to identify the root cause of the issue

In theory, this seems straightforward. So why is this flexibility so difficult to achieve in reality?

Data Warehousing and the Data Analyst Feedback Loop

For nearly 30 years, data warehousing has been the standard model to aggregate data and provide business-directed analytics. Data is extracted from various sources, transformed to a predefined model, and loaded into the data warehouse. This extract, transform, and load (ETL) process results in queryable analytics contextualized by key dimensions (e.g., customer, product, location). But this process is slow and leads to latencies in the data warehouse. Stale data (even just a week or two old) can be useless data for many purposes.

Metrics defined from data warehouses can be too broad or inflexible to guide nimble decision making. This limits your ability to drill down into the source data or investigate the data from a new perspective, which doesn't make the data actionable.

In the traditional model, it's not atypical for analysts to use multiple Excel spreadsheets, a transactional database, supplementary databases, and an enterprise resource planning (ERP) solution to guide their reports. Analysts' independence in using various sources and tools can lead to issues with consistency and accuracy. Without a consistent source of truth for data definitions, confusion and errors can result.

NOTE

Consistency: Are complex analyses (affinity analysis, multi-criteria decision analysis) calculated the same way?

Accuracy: How do you ensure accuracy of the data and analysis between various analysts?

Furthermore, data becomes siloed inside of the data warehouse, which restricts analysts' abilities to access necessary data quickly. Analysts can get stuck in a loop of user requests, custom reports, and Structured Query Language (SQL) queries, while the decision makers are limited to asking a few questions at a time.

Though each of these issues presents a challenge, the overarching problem is that the data is separate from the action. Data centralization alone is not the answer—it must go hand-in-hand with a data-driven culture.

The Impact of the Traditional Data Warehouse Model

- ETLing data into a data warehouse can be slow, leading to stale insights
- Metrics can be too broad or inflexible, preventing nimble analyses
- Data silos make analysts report generators and query writers
- Lacking a "single source of truth" can lead to issues with definitions and accuracy

Building a Data-Driven Culture

What Does It Mean to be Data-Driven?

Carl Anderson, the Director of Data Science at Warby Parker, outlines these six characteristics of a data-driven organization.[3] Such an organization:

- Is continuously testing
- Has a continuous improvement mindset
- Is involved in predictive modeling and model improvement
- Chooses among actions using a suite of weighted variables
- Has a culture where decision makers take notice of key findings, trust them, and act upon them
- Uses data to help inform and influence strategy

How Can Data Centralization Contribute to Becoming Data-Driven?

- The emergent properties of centralized data allow for a company to quickly act upon new findings
- Consistent definitions (a single version of truth) build trust in the analytics (which makes it easier to act upon them)
- Avoiding the data breadline/bottleneck frees up key team members to investigate new inquiries and perspectives

What's the ROI?

Considering the hype and complexity of a centralized data system, it's important to ask if there is a tangible ROI for this type of investment. A Nucleus Research report found that in 2011, there was a 10.66:1 return on investments in analytics.[4] In 2014, Nucleus found that return increased to 13.01:1.[5]

How did the usage of new analytics tools lead to this ROI? Nucleus proposes that the decreased complexity to integrate data sources

3 Carl Anderson. *Creating a Data-Driven Organization*. Sebastopol, CA: O'Reilly Media, 2015.

4 "Analytics Pays Back $10.66 for Every Dollar Spent." *Nucleus Research*, December 2011.

5 "Analytics Pays Back $13.01 for Every Dollar Spent." *Nucleus Research*, September 2014.

with analytics applications eliminated manual processes for report builders and SQL writers. Analytics enabled better decisions with a significant increase in profitability. They also found that the benefits were not limited to expert application users (meaning a company wouldn't have to invest in personnel expertise in addition to purchasing the tool), nor to a particular sector or company size.[5]

With a nod to data centralization, Nucleus also found that the highest ROI resulted from departments that made data more available to decision makers, and that integrating the analytics application with three or more data sources achieved higher returns.

ROI of Integrated Analytics

In 2011, every dollar invested in analytics paid out $10.66. In 2014, the ROI increased to $13.01.

The ROI was not limited to expert users or particular sectors, and increased when analytics tools were integrated with three or more data sources.

Roadmap to Data Centralization

The path to centralization will vary based on the types of data, size of the company, and needed metrics. But we will begin with the human element—becoming data-centric relies on stakeholders identifying and agreeing upon an approach, definitions (a source of truth), and the data pipeline.

The Argument for Data Centralization

For disparate data sources to be compared, they must contain common fields that can be mapped or linked. Evaluate each data source for existing common fields and, if you can, resolve minor variances (for example, region vs. state vs. zip code). You could also standardize data references, though some tools will allow you to specify relationships without needing to unify labels (e.g., product_id vs. product_number).

SaaS data streams can be particularly difficult to link, as many use unique fields that can be difficult to identify and unify across multiple products. If you don't have in-house expertise, data intermediary

or integration tools (like Fivetran) can pipe SaaS data streams into a data warehouse that will play nicely with a variety of analytics tools. These intermediaries could also help you upgrade to next-gen databases (like Redshift, Vertica, and Snowflake), which may expand your capabilities when you select your company's BI tool.

Identify Stakeholders

Going back to one of Carl Anderson's characteristics, decision makers in a data-driven organization take notice of key findings, trust them, and act upon them. Building a culture of trust and awareness requires a collaboration between decision makers, data analysts, and quality management.

Key Players and Functions for Building a Data-Driven Organization

Decision Makers

- Define the business needs (specify metrics)
- Support the data-centric initiative
- Institute and encourage training/accessibility to new tools
- Act on the analytic findings
- Provide feedback on how the analytics affected decisions

Data Analysts

- Evaluate the analytic product(s)
- Identify expertise gaps
- Define source data streams
- Create and agree upon key definitions (sources of truth)
- Request feedback, then iterate on analyses

Quality Management

- Define a data governance policy
- Create a data classification hierarchy
- Specify access restrictions and permissions according to defined policies and procedures

Create a Data Plan

With the team in place, create a data plan.

Step 1. Define needs and specify your metrics. What key metrics impact decision making (sales, profit, users, customer happiness)?

Step 2. Define measurements. Can these metrics (e.g., profit) be measured directly? If so, from what data streams? If not, what data should be used to correlate with the key metric (for example, what would be used to measure customer happiness)?

Step 3a. Identify data sources (master data). Where is your data coming from?

- Databases
- Data warehouses

- SaaS/Cloud products (Marketo, Facebook, Salesforce, Zendesk, website analytics)
- Product event tracking
- Public data sources (census data, scientific data)

Step 3b. Identify gaps. What's missing? If you find that a key metric isn't measured, how could it be measured? Do you need any additional expertise or consulting to achieve this plan?

Step 4. Prioritize. You may not be in a position to centralize all your data right away. Prioritize centralization for your most important metrics, and pick tools that will allow you to centralize additional sources over time.

Step 5. Standardize your definitions. Create a single source of truth for analyses. Some metrics—like sale, profit, or user—may be simpler to use consistently. More complex or subjective analyses, like affinity analysis or multi-criteria decision analysis, provide more value when standardized across an organization.

Step 6. Data governance. Increasing access to a centralized data resource poses a risk. If you don't already have a data classification policy in place, now is the time to create one. Consider the data streams you identified above—can you classify all the data provided by each stream? What access restrictions should be in place, and how should those restrictions be controlled (by user or team)?

Step 7. Evaluate accessibility. Who from the organization (persons and teams) should have access to the centralized data? How will you ensure that they have access? How will you provide training and support?

Once the plan is defined, bring in key members of each team or department. How will this data impact their day-to-day? What other perspectives or data streams would be useful?

Find the Right Tool(s) for the Job

While a comprehensive review of all BI tools on the market would exceed the scope of this report, we can categorize these products to help you find the right tool for the job.

Legacy architecture tools

Enterprise tools such as IBM Cognos, Microstrategy, Oracle BI, and SAP Business Objects (among others) create one large data model

around all of your data streams. If a key element of your data pipeline (a data warehouse, ERP, or customer relationship management [CRM]) is from an enterprise vendor, you may consider a BI solution from the same source (e.g., Oracle database with Oracle BI). This could allow you to directly (or more easily) integrate data from those key products. The familiarity of a particular vendor could increase adoption of a new BI/Analytics interface.

Qlik offers an in-memory, more modern take on legacy BI platforms with QlikView, though its self-service capabilities are limited. In response to visualization-heavy platforms like Tableau, Qlik now offers QlikSense, a self-service visualization interface.

When calculating the cost of any new tool, include price-per-user (if licenses are used), the expense of implementation resources, and any in-house resources needed for maintenance (in addition to the cost of ownership). Enterprise or legacy architecture tools can take longer to implement and be more expensive to deploy. If you have next-gen databases or cloud data streams, you'll also need to evaluate how well these tools integrate that data.

NOTE

Legacy tools build one large data model around all of your data streams

Advantages:

- May work best on data warehouses or other key products (ERP, CRM) from the same vendor
- You can select from on-disk or in-memory processing capabilities
- Vendor recognition/familiarity can increase adoption

Disadvantages:

- May be expensive and time-consuming to deploy
- May be difficult to utilize next-gen databases or cloud data streams
- Need to ensure this offers flexible, nimble access to the data and drill-down capability (not just a vanity dashboard)

Data visualization platforms

Data visualization products, like Tableau, offer a self-service analytics tool built on pre-defined data tables. Users create their own reports with this data through a visual, drag-and-drop interface by selecting dimensions (values) and measures (counts, sums, averages, etc).

End users can also execute joint operations or integrate ("blend") data from different tables. Using VizQL, the program translates drag-and-drop actions into SQL queries. The interface also allows you to drill down through data hierarchies and into the source data tables.[6]

In the Data Server, analysts or database (DB) experts can join data streams, rename fields, create an alias for particular values (like null), or create measures. Tableau can utilize SQL databases, Hadoop, Redshift, and others, and Tableau Online (Tableau's hosted, SaaS offering) also supports cloud connections like Google Analytics.

The intuitive, self-service interface and the clean graphics of a visualization-forward tool makes this platform attractive to many users. The most significant disadvantage is that a pre-organization step is required to create the data tables accessed by end users. This limits data exploration and drill-down, as end users do not have full flexibility to manipulate the data (they are unable to edit the data model or data hierarchies).[7]

In the end, however, users are still accessing only some (not all) of the data, resulting in a potentially incomplete picture. Adding new data or evaluating data from a new perspective requires an information technology (IT) team to update the defined/pre-organized data table(s). This workflow can lead to the same limitations as the traditional ETL model, such as limited flexibility for analytic insights, data bottlenecks, and the lack of a single source of truth.

6 "Drill Down and Hierarchies." *Tableau Video Tutorial*, 2 December 2015.

7 "Data Server." *Tableau Video Tutorial*, 2 December 2015.

NOTE

Visualization tools provide a self-service analytics interface

Advantages:

- Intuitive, drag-and-drop, self-service interface
- Data connections can pull real-time or near real-time data
- Polished graphics from a visualization-first product

Disadvantages:

- End users work on top of a predefined data table, which must be built ahead of time by experienced users
- Pre-organization workflow results in limited flexibility and data exploration
- Cloud or SaaS data streams may be difficult to integrate

Managed and cloud-only services

If your data pipeline primarily exists of cloud data streams, a modern approach (cloud-based service) may integrate your data more easily. Cloud-based solutions are more friendly to SaaS data sources—sometimes even offering a direct connector that can ease the complexity of centralization.

If you don't have strong in-house expertise, managed cloud-based services may also be a good choice. These solutions will pull and centralize all of your data for you (typically, using a massively parallel processing [MPP] database). For example, Domo is built on top of an HP Vertica warehouse infrastructure[8] and GoodData uses an MPP data warehouse on Amazon Web Services.[9]

With a fully managed model, you may lose transparency on how the data is joined and analyzed. If eventually you want to take this in-house, can you replicate the managed process to ensure your definitions or key metrics won't change?

8 Mark Smith and Jeff Morris, "10 Reasons Why Smart Organizations are Moving to Cloud BI." *GoodData Webinar*, 2 December 2015.

9 Matt Aslett, "Domo Emerges From Stealth with Cloud-Based Business Management Platform." *451 Research*, 7 May 2015.

With the difficulty of SaaS streams, it's also possible that the managed service is unable to centralize everything, meaning that time and resources have been invested into a model or service that may not deliver.

Since the data will be stored by a third party, security is a common question for managed services. Each vendor's security practices should be considered according to your data classification policy. Some providers, like GoodData, are Health Insurance Portability and Accountability Act (HIPAA) compliant and offer in-transmission and at-rest encryption for sensitive data.[7]

NOTE

Cloud-based tools allow you to outsource your analytics

Advantages:

- Friendlier for SaaS data streams
- Managed services provide assistance without expensive hires or consulting services

Disadvantages:

- You may lose transparency on how the data is joined
- Since the data is now managed by a third party, you should evaluate their security according to your data classification policy

Data exploration platforms

The newest analytics products offer an all-in-one exploration platform. Looker is a business intelligence startup whose product is built with an extensive modeling language (LookML) and a self-service interface. This allows users to build and share their own visualizations (just like a dashboard), but also drill down and explore the entire database in full detail without any SQL queries.

Looker leverages the increased processing power of next-gen databases and integrates with a variety of compatible data sources (SQL databases and Hadoop via SQL interfaces like Cloudera Impala, Hive, Pivotal HAWQ, and Spark).[10]

10 "Analytics Everyone Loves." *Looker*, 2 December 2015.

The Looker interface allows for an "in-database" or "schema-on-read" model that breaks from the ETL paradigm. There's no need to design or maintain a resource-intensive transformation step to format your data; this is all done natively within a modeling layer.

Integrating cloud sources can represent a difficult task for any BI platform. You can create API calls to dump data into a centralized database ahead of time. If you take care of the E&L part of ELT, Looker allows you to transform the data in their modeling layer from there. Recently, Looker introduced templatized code, "Looker Blocks," to decrease the complexity of integrating third-party data sources, like Salesforce, Zendesk, and Marketo ("Source Blocks").

Once a connection to a particular data source is made, Looker creates an editable data model that finds and describes the tables and relationships within that database. Looker's modeling language, LookML, is built on top of root sources (column names, table names, data types) to create the dimensions and measures that will be used to explore your data. Technical users can build complex measures from SQL queries across multiple data sources. These measures can be saved, referenced, and re-used as a shortcut by non-technical users. Analysis Blocks (another type of Looker Block) templatize analytics that traditionally require difficult SQL queries, allowing plug-and-play analytics by non-technical users.

Using a simple drag-and-drop interface, users can select data models of interest, filter or pivot the data as needed, and build visualizations like a dashboard. If a user identifies a particular dip or spike they want to investigate, they can drill down into the source data and further analyze that isolated point by slicing it with other data (all without writing any SQL queries). As you explore, you can switch back and forth between the visualization, the table, and the SQL that Looker is writing for you.[11] This structure allows for a fully reversible flow path from source data to visualization and back again.

11 Looker Demo, "From Database to Dashboard." *Vimeo*, 2 December 2015.

NOTE

Data exploration platforms provide a visualization and in-database query tool

Advantages:

- In-database query allows for flexible exploration without the need for predefined tables
- Data models create a single source of truth for users
- Performs best on next-gen databases
- SaaS data integration is still challenging, but improved with Looker Source Blocks

Disadvantages:

- Visualizations may not be as polished as a visualization-first product
- Experienced personnel may need to assist with the E&L steps of ELT, but transformations can be done from there within the modeling layer
- Performance won't be optimized on older databases

Plan Forward

Whichever tool you choose, it's important to go back to the theory of actionable data to ensure the product is meeting your needs as a data-driven organization:

- Does it restrict data to analysts, or does it enable self-service investigation?
- Can users ask any new questions on the fly from any data source?
- Does it only provide slow, broad metrics, or does it allow for real-time, actionable data insights?
- Does it provide a vanity dashboard, or can you drill down into source data to investigate a new question or perspective?
- Does it house and maintain single sources of truth?
- Can you easily integrate new data sources and take advantage of next-gen tools?
- Is it usable and accessible enough to encourage adoption within the entire organization?

Conclusion

Centralizing data streams provides an emergent property (context) that enables better business decisions. The operational side of centralizing data is a difficult task, but the process of defining key metrics and single sources of truth will pay out over time (regardless of which vendor or product you choose).

It's not enough, however, to just centralize your data or to just define a data-driven culture for your company. The idea and the tools must go together to make data accessible, flexible, and actionable.