DIYOTTA

Modern Data Integration

White Paper

DiYOTTA

Table of contents

DiYOTTA

Challenging All Conventions

Preface

Advances in technology continue to accelerate the pace and competitive environment of business. Those organizations that are able to utilize information to analyze activities and trends, and create new insights are leading their industries. Business leaders rely on fact based decision-making and information analysis for their competitive advantage. Innovations in technology ranging from automation of manual activities to the interconnectivity of devices for the Internet of Things are contributing to the overwhelming amounts of data.

Data complexity has also grown due to the increased number of data sources, new data types, greater locations where the data resides, and increased volumes of data being generated. Welcome to the era of modern data…not just big data. Organizations have to wrangle modern data in order to have greater visibility into their operations, customers, markets, competitors, compliance with regulations and to create actionable insights in a timely manner just to stay relevant.

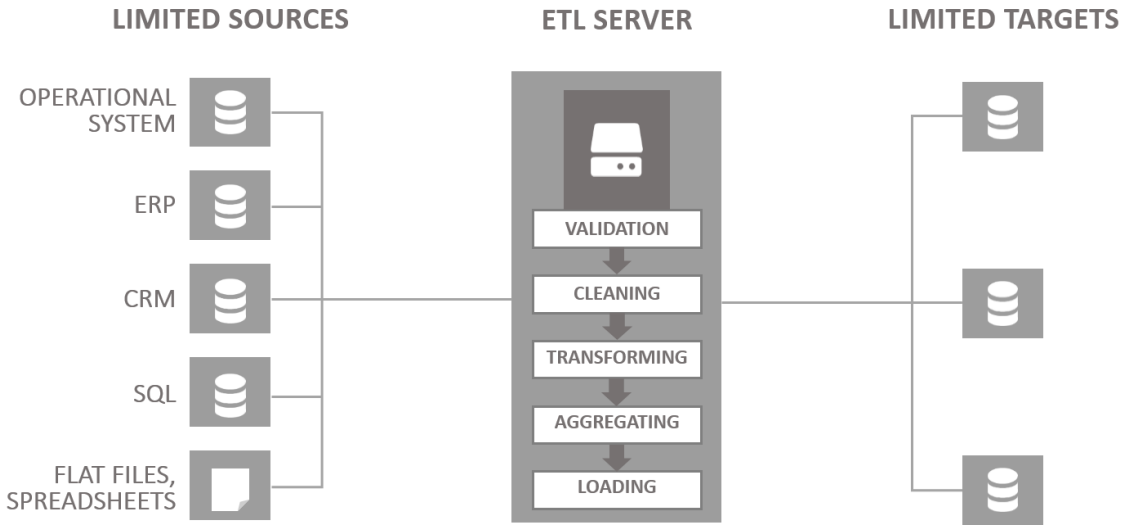
Data storage and processing platforms, such as Hadoop, were created to address modern and big data. However, traditional data integration technologies that extract, transform and load (ETL) data hinder the speed of data availability and in many instances, do not work with modern data due to their architecture. The process of moving data from source to target through a data integration application server causes big data bottlenecks. Many IT professionals are forced to abandon traditional data integration applications for custom coded scripts. While custom code can solve the throughput problem, it creates a whole new set of problems associated with maintenance, documentation, and knowledge transference. Organizations are demanding greater information agility, insight, and value from their investments in modern and big data. A new and more modern approach to data integration is needed.

The principles of modern data integration were created out of the frustrations brought on by working with traditional data integration technologies and having to maintain custom coded scripts in a big data environment. The founders of Diyotta, with their vast technical experience, challenged conventional thinking about data integration and have created the principles of modern data integration that are needed in today's modern and big data environment.

The Paradigm Shift

There is a major paradigm shift taking place in the field of data and information management. Every day, there are massive amounts of new data being created. Unlike ever before, this data is moving fluidly in many directions; from on-premises to the cloud, cloud to on premises, and people are accessing data from all over the place. The old paradigms of data integration have become unsuitable in this new data landscape. It is time for a paradigm shift. Now is the time to make the move to modern data integration.

Think about where we've come in the last twenty years. When data integration software was just emerging on the scene in the early-to-mid nineties, the sources for data were limited. Most of the data came from mainframes, operational systems that were built on relational databases, and data service providers. The targets were primarily data warehouses and data marts, with an occasional loop back to operational systems for closed loop Business Intelligence. It is no surprise that most of the prevailing data integration software we use today were created in a vastly different world with very different demands.



Most data integration software was built to run data through ETL servers, as depicted in the above diagram. It worked well at the time for several reasons: there wasn't that much data – 1TB was considered a large amount of data; most data was structured, and the turnaround time for that data was monthly. Even back then, daily loads became a problem for most companies. Because of the limitations of the early data integration software, much of the work was custom coded, without documentation, and no central management. If these legacy data integration technologies had grave challenges back then, consider how much more obsolete they are in today's modern and big data world.

The Shift in Data

We are already well into the age of big data with more data than ever before, "From 2013 to 2020, the digital universe will grow by a factor of 10 – from 4.4 trillion gigabytes to 44 trillion. It more than doubles every two years."[1]

New data and new data types are emerging every day with very limited structure. In fact, 80% of all enterprise data is unstructured or semi-structured.[2]

Data is also coming from many different places. There are a growing number of mobile devices and applications producing data**;** 20 billion devices are already connected to the Internet and by 2020, this number will grow by 50% to 30 billion connected devices.[3]

Data in SaaS applications and the cloud will also continue to grow at record pace. In 2013, about 20% of the data in the digital universe was "touched" by the cloud; either stored, perhaps temporarily, or processed in some way. By 2020, that percentage will double to 40%.[4]

Data is pouring out of thousands of new API's that are created by apps and the Internet of Things. There are billions of sensors and devices creating trillions of time stamped events and the number of sensors will soon number in the trillions. "Fed by sensors soon to number in the trillions, working with intelligent systems in the billions, and involving millions of applications, the Internet of Things will drive new consumer and business behaviour that will demand increasingly intelligent industry solutions…"[5]

---

[1] EMC Digital Universe Study, 2014 – Executive Summary (http://www.emc.com/leadership/digitaluniverse/2014iview/executive-summary.htm)

[2] According to IDC, Forrester and Gartner (http://www.eweek.com/storage/slideshows/managing-massive-unstructured-data-troves-10-best-practices#sthash.KAbEigHX.dpuf)

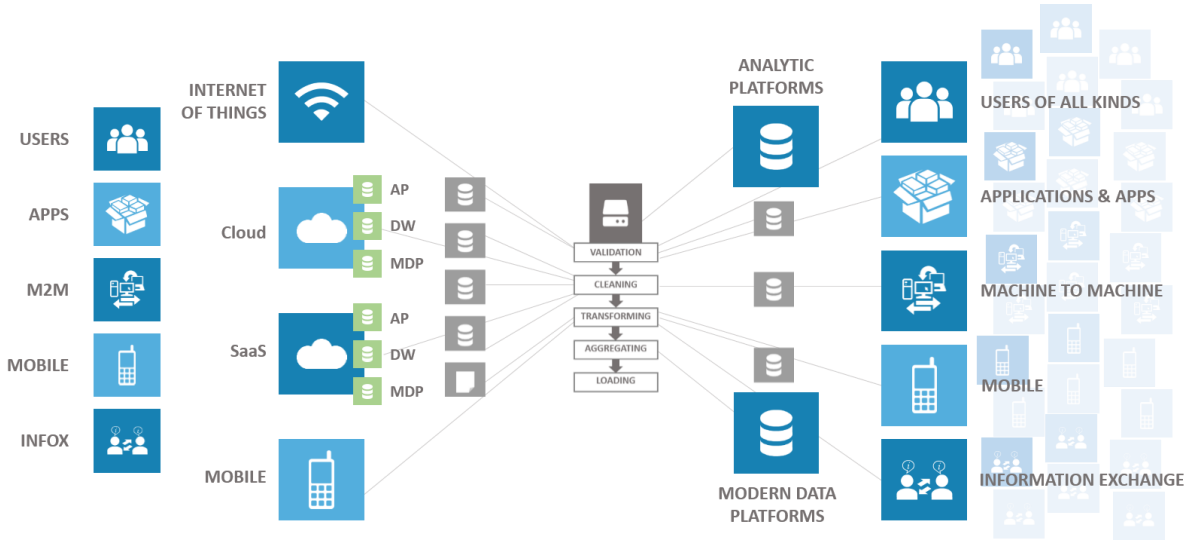
[3] Ibid.

[4] EMC Digital Universe Study, 2014 – Executive Summary (http://www.emc.com/leadership/digitaluniverse/2014iview/executive-summary.htm)

[5] EMC Digital Universe Study, 2014 – Internet of Things (http://www.emc.com/leadership/digitaluniverse/2014iview/internet-of-things.htm)

The Shift in Complexity

To further complicate things, the number of platforms and targets where data is being landed has also grown to include many new open source and purpose-built platforms. In addition, there has been a rapid expansion in information consumers. The mix of data users now includes all classes of data users and analysts, plus applications, machines, mobile users, and massive communities of data-savvy end users. What used to be hundreds and thousands of users has now grown to hundreds of thousands of users, and in some cases, millions of users.



Complexity continues to grow with data that used to be stored on premise, now also being stored in the cloud and in SaaS environments. Information consumers that were previously confined to data on premise are also now accessing data in the cloud and SaaS data stores. Data that used to flow in one direction now flows in all directions. We now live in a matrix of data where the complexity of sources and targets continues to grow incrementally and there is no end in sight, as depicted above.

New Challenges Require New Approaches

Why use a film camera to take a photo when you can use digital? While camera film technology works, it is outdated and expensive. That's exactly what is happening with companies who are trying to use legacy ETL technology to process and provision data in today's interwoven world of data.

An ETL server becomes a huge bottleneck; flowing everything through a single server image creates massive overuse and network congestion. Many of the legacy data integration tools have been re-engineered to handle new kinds of data, but there are only so many add-ons you can try to attach to old technology. Multiple connectors and modules bring even more complexity into an already complex problem. Trying to bring an "old world" tool into this new landscape requires massive amounts of custom coding for the growing mix of data types, sources, and API's. Thus, code management continues to be a nightmare for those making the move to modern and big data.

Meanwhile, lack of documentation and limited reusability means that the work has to be redone every time new data is added or new platforms are implemented. Ultimately, this amounts to more money spent, more time and resources wasted, and less effort spent on bottom line efforts and pure innovation. Now more than ever before, new challenges require new approaches.

Big Data Changes Everything

This discussion wouldn't be complete without factoring in the impact of big data and Hadoop. Hadoop adoption is skyrocketing, driven by three main drivers:

- It solves technology challenges like scalability, performance, and maintainability.
- It empowers the business to forge new ground in both discovery and predictive analytics.
- It is affordable enough to allow companies to keep all of their data and discover the value later.

While some still question the longevity of the Hadoop explosion, at this point, we have passed the tipping point. Hadoop is here to stay and gaining more ground every day. The pace at which we are moving is no longer linear, but exponential; so now is the time to embrace all that we can do with big data.

Hadoop must also live within the existing corporate and Internet environment. However, it is not the be-all and end-all of data platforms. There are many things that it does well, but there are also many things that are done better on other kinds of platforms—like analytic platforms. For this reason, Gartner has come up with the Logical Data Warehouse and the Context-Aware Data Warehouse. Essentially, multiple platforms must live and work together.

Even though Hadoop may end up being a big player in big data, today it is one of many players. As a result, some of the greatest challenges involve moving data into and out of Hadoop. Hadoop is both a landing spot for massive data coming in and a provisioning platform for processed data coming out.

Five Principles of Modern Data Integration

We've established that we live in a new era of data. It's not just big data, it is modern data with new types, sources, volumes, and locations of data that we have never before experienced. We also live an era with new solutions emerging to meet the most pressing data integration challenges.

Modern data is more complex and more distributed than anything we have encountered in the past, and if we are going to take advantage of it, then we must connect emerging data with modern data integration. It's time for a shift in paradigm so that we can spend more energy gaining insight from our data rather than wrangling and fumbling around with it.

There are five key principles of modern data integration that unlock unprecedented new insight from the matrix of data that surrounds us. Working together, they take advantage of the emergence of new data and new platforms, rather than fighting against the rising tide.

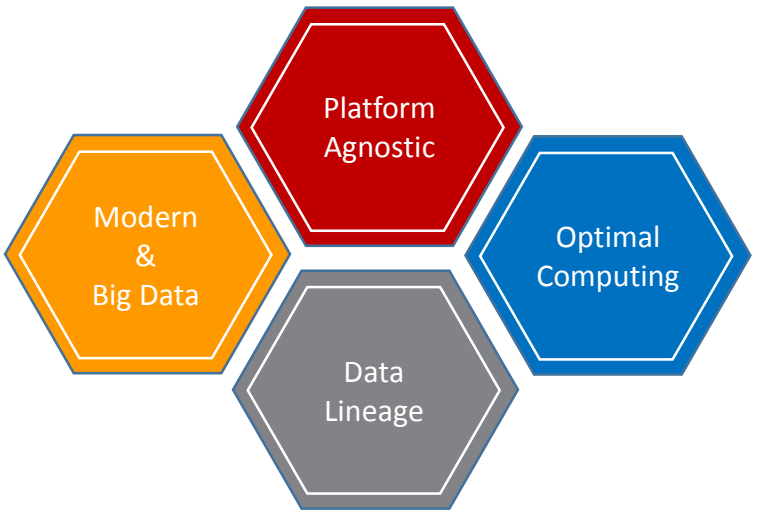
- **Take the processing to where the data lives.**
  There is too much data to move every time you need to blend or transform it. It makes more sense to place agents near where the data lives and process it locally. Carefully coordinated instructions can be sent to the agent, and the work is done on the host platform before any data is moved. By taking the processing to where the data lives, you eliminate the bottleneck of the ETL server and decrease the movement of data across the network.

- **Fully leverage all platforms based on what they were designed to do well.**
  You've already invested in powerful databases and you are making new investments in modern data platforms. Every platform has a set of workloads that it handles well and a set of built-in functions. Modern Data Integration allows you to call those native functions locally, process them within powerful platforms, and distribute the workload in the most efficient way possible. By fully leveraging existing platforms, you increase the performance of all of your data blending and enrichment while minimizing data movement.

- **Move data point-to-point to avoid single server bottlenecks.**
  Since data is moving in all directions, it is vital to establish processes that move data point-to-point, rather than through a data integration server. This new approach gives you the ability to move data at the right time. There are times when you want to move all of your data; but most of the time, it makes more sense to process it in its native environment, and then move the result set. By moving data point-to-point and eliminating server bottlenecks, you get massive network savings and increase the speed at which you transfer data.

- **Manage all of the business rules and data logic centrally.**
  Operating dispersed integration applications creates chaos in today's changing data ecosystem. New data and modern data platforms must be managed centrally with all business rules and data logic in a single design studio. You can only do this in an architecture where a central design studio is completely separate from local processing agents using native functions. Managing all of your business rules and data logic centrally gives you complete transparency, accessible lineage, and maximum reuse. You'll never have to waste your time redesigning flows.

- **Make changes using existing rules and logic.**
  With all management handled centrally, you are also able to keep all the existing business rules and data logic templates in the metadata repository. As a result, when changes need to be made with new data, new platforms or even migrations from one platform to another; you are able to make those changes quickly, using the existing rules and logic.



Five Capabilities of Modern Data Integration

When you make the move to modern data integration, you will eliminate many of the challenges brought on by legacy data integration technology. In addition, the new paradigm brings with it five new capabilities, well beyond anything you have experienced with traditional data integration technology:

- **Design once and use many times.**
  Reuse is not a new concept, but the level of reuse possible after modernization will be significantly increased through the conversion of native functions from one platform to another. For instance, if you are moving data storage and processing platforms from Oracle to Netezza today, and then sometime in the future you want to move from Netezza to Hadoop, you already have at least 80% of all of that business rules and data logic stored for reuse.

- **Gain complete knowledge of your data**.
  In modern data integration, much of the metadata is generated based on what you read out of the source platform. You'll capture the remainder of the metadata as you go through the process of designing your blending and enriching. Ultimately, you will know everything you need to know about your data and have it all stored in one central location.

- **Manage complex data environments without complexity.**
  By storing your logic centrally, deploying agents locally, and sending out instructions as needed, you remove the complexity of data flowing in all directions through many different servers. As your environment become more complex and lives in hybrid environments, you'll be able to deploy multiple agents to wherever the data lives while operating all of your data from one central location, in one consistent manner.

- **Optimize your actions.**
  When you are moving data from one platform to another, you need to be looking at the complete set of transformations that need to happen and assign them to the highest performing platform. This allows you to optimize performance across your entire data matrix. For example, you can send your analytic workloads to an analytic platform and your text processing to Hadoop.

- **Adapt quickly.**
  Business does not stand still and all businesses require action. Having the ability to change, extend and migrate existing business logic and data flow designs into future is, perhaps, the most valuable aspect of modernization. A typical business logic change request often meets the response, "That's nice. Let's wait six months." One of the great things about modern data integration is being able to respond quickly to those business needs.

Modern Data Integration in Action

Scotiabank

Scotiabank, a large international bank headquartered in Canada with a footprint across the globe— 25 million customers in 55 countries, 86K employees, 22 billion in revenue, 7.3 billion in net worth—wanted to improve customer insight by expanding their data warehouse capabilities. They were tired of the high cost of their traditional warehouses and the inflexibility in those platforms to deal with large volumes of historical data, different variety of data and ability to get intraday data from various parts of the organization including mortgage, real estate, securities and line of credit so they made a decision to switch to Hadoop platform.

Using Diyotta Modern Data Integration Suite, the entire design process for the migration was completed within two days, all data and definitions were moved into Hadoop in under a week and they are now moving over 200 tables with 10GB of data daily delivering valuable insights to their business users for further analysis

Originally, this was a migration project they thought was going to take six months. But with Diyotta Modern Data Integration Suite, they got it done in just three weeks. They're using 100% of their existing resources and they were able to lower costs by more than 50% compared to the other competitive ETL offerings. Beyond that, they are able to reuse 80% of what they did in this migration for future migrations. So, if they want to move to Spark sometime in the future, they will be able to reuse all that they've done.

Sprint

Sprint, one of the largest wireless service providers in the United States and a major global internet carrier, wanted to have greater visibility into their cellular network. With 59 million subscribers generating 12 billion records or 6 terabytes of data a day, Sprint was looking to create a solution to collect all of their call data records from over 100 data feeds and then provision the data from Hadoop to an analytical application for fraud detection.

Diyotta Modern Data Integration Suite was selected to integrate the data sets into Hadoop in near real-time. Sprint was able to complete the project three times faster and realized 75% cost savings by using Diyotta Modern Data Integration Suite instead of having an outside vendor create custom code. The cost savings was in excess of $1.4 million. To learn more about this case study, you can watch a video of Tim Connor, Manager, Center of Excellence, Big data, Advanced Analytics, and Data Science at Sprint by visiting Diyotta.com and then by searching for "Sprint".

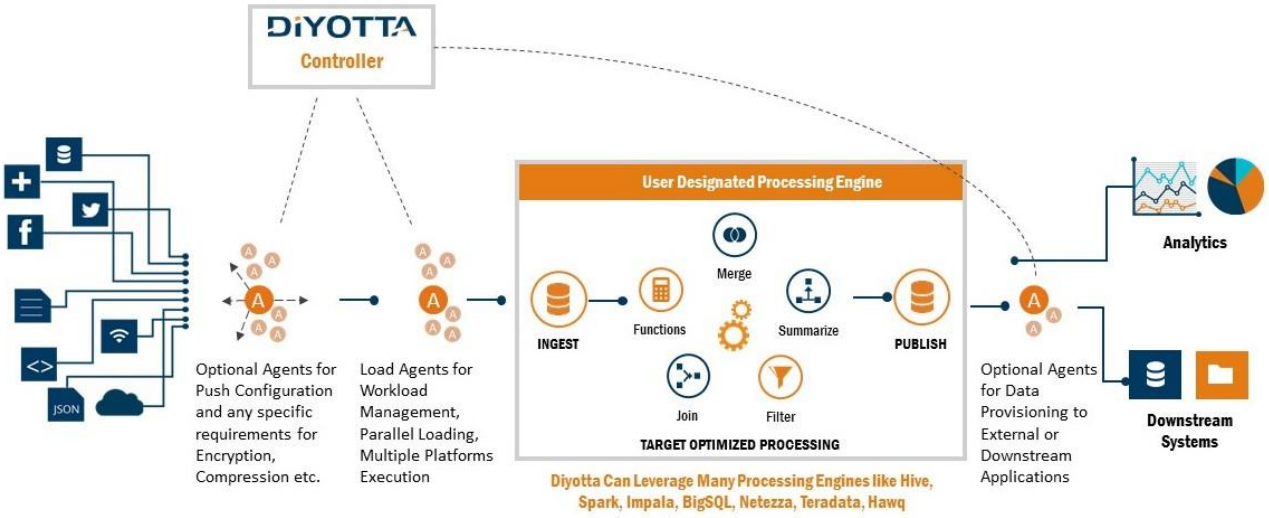
Looking Beyond Modernization

Think about what lies ahead on this road to unleashing new paradigms. The emerging data landscape is only accelerating. The transition to stream data on-demand, in real-time will soon become the norm in business. Businesses currently looking at monthly reports will be looking at daily reports; daily reports will become hourly reports and even minutes and seconds.

As an executive, it will be essential to leverage real-time data to know exactly what is happening in your business right now. You need to know anything that might damage or benefit your business immediately. Where fraudulent activities are happening in the financial sector? Are your customers happy with your products / services or dissatisfied? How are buying patterns affecting your business? Processing all that data in real-time will only grow in necessity. You need the quickest time to value outcome, transparency in your data, and adaptability to meet these needs.

Technology may be changing, but the requirements are not. You need to be able to design your data movement, blending and transformation with complete transparency and unlimited reuse. You must be able to accelerate the execution of ongoing data enrichment and analysis by deploying agents that deliver instructions and optimize actions. It is essential that you adapt your environments to meet the ever-changing business and technical demands; migrating data from platform to platform should be effortless in these changing times.

Now that you know the values, how will modern data integration influence the most pressing projects in your enterprise? We've built a tool that encompasses all five principles and capabilities of modern data integration. Diyotta helps you leverage what is happening across all of your platforms, turning big data into an information hub. We are the industry's first data provisioning platform, purpose-built for big data—hybrid system and the cloud—by the industries top data integration experts. Only we offer a design-once, multiple use architecture for rapid change of both data and platforms. We've built a tool that encompasses all five principles and capabilities of modern data integration, as depicted below.



Let's get started by arranging a deep dive of our offerings and letting us know about your information environment. Take control of your data with Diyotta.

3700 Arco Corporate Drive, Suite #410, Charlotte, NC 28273
+1-704-817-4646 | +1-888-365-4230 | +1-877-813-1846