Big data lakes often acquire a large amount of unstructured data, plus structured ungoverned data. To discover meaningful insights, the data must go through significant iterations of discovery and visualization by data scientists. Without a proper framework, these iterations can be repeatedly attempted in inefficient ways. In this article, explore a pattern for data lake exploration and learn the steps for data integration across the end-to-end data flow. Through a series of exploration use cases, the tutorial defines the characteristics of the exploration and their execution by using the pattern.

View more content in this series
Introduction

The previous tutorial in this series described the predictive analytics in the Advanced Analytics Platform (AAP). It discussed how to use predictive analytics to determine mobility profile patterns and text analytics to gain insights into text data. Such analytics is possible because of the Advanced Analytics Platform, which integrates the different software components together.

The ability to discover and visualize data can enable an enterprise to interpret and analyze its data much more effectively. For example, with a complete understanding of your customers, you can provide them better products and services. To do so, you need to understand who uses the product, how the product is used, and future trends for the product. To achieve this understanding, you need to discover the appropriate data from large amounts of information, enrich the data by aggregating the information from multiple data sources, and then visualize the data. The appropriate data might, for example, include CRM and Social Media data about the customers. Enriching the data would include integrating the social media data with CRM data to obtain a richer customer profile. This data is then aggregated with more sources such as sales, marketing, and product information. This integrated data can then be visualized by using key performance indicators, interactive displays, and relationship graphs. This process is developed in multiple steps and repeated to improve the visualization. This pattern is repeatable, and can be iterated multiple times to extract the correct data source.

Unlike carefully governed internal data, most big data comes from sources outside your control and therefore suffers from significant correctness or alignment problems. To make meaningful insights, you must judge the source credibility and source definition diversity. Veracity represents both the credibility of the data source and source definition diversity of the data. The exploration must account for veracity and provide adequate process to align, prioritize, and integrate the data before you present the results to a data user. We call the pattern that enables the preceding exploration and visualization the data lake pattern.

The associated tasks for the data lake pattern can be assembled into a framework with the following components:

- **Discovery**
  - **Business understanding:** This task includes gathering the business use of the data, the requirements for the modeling of the data, and the business impact of the data.
  - **Data modeling:** Create a conceptual model of the elements that are required to answer the business requirements. This step includes the creation of the ontology and the architectural context.
  - **Data inventory:** The initial step in the creation of a data audience is to review all the possible sources of information and catalog them. The creation of an ontology helps to document the semantics of the data objects selected.
  - **Data selection:** Determine the multiple heterogeneous data sources that might contain the data for extraction. In this step, one determines which specific data sources are required to create a complete picture of the entity or event that is being analyzed.

- **Integration**
• **Data unification:** The combining of all the sources of data to match the conceptual model that is created for the business problem. One creates access to all data sources by federating them, either creating routes for the data or integrating their content.

• **Data aggregation:** Determining the level of aggregation of the data that is required for the problem, setting a level of aggregation that responds to the requirements in the business problem. The aggregation level can correspond to multiple dimensions.

• **Data transformation:** After defining all the metrics to create homogeneous data from the heterogeneous data sources, you apply data transformations on the sources to create a homogeneous data set that responds to the business problem.

• **Entity analytics and resolution:** The source data can use different identifications. In addition, the data might suffer from data quality issues that result in duplicate or ambiguous relationships. With entity analytics, you can align the data by using a set of rules and hidden relationships across the data. You can use sophisticated statistical techniques to establish the degree of fit and rules to either automatically align the data or use user assistance to resolve differences.

• **Data mining:** After all the data is collected, you analyze the collected data to determine patterns in the data that might be relevant for the problem addressed.

• **Visualization**
  
  • **Definition:** Establish a glossary of terms, user-defined dimensions, measures, and relationships. These definitions are used for data exploration by users.

  • **Ad hoc queries:** The most natural form of data exploration is in the form of open-ended queries. These queries allow users to view a sub set or aggregation of data by using their query parameters. The queries can be generated by using GUI tools or SQL-like commands.

  • **Geo/semantic mapping:** The creation of an integrated visualization to respond to the business requirements by using either a geographic or semantic model for representation. In the geographic model, the data is overlaid on a geographic map. In semantic mapping, the data is overlaid on a user-defined model, such as organizational hierarchy or customer segments. This task includes connecting the data representation to the visualization components, which might be spread across multiple dimensions.

  • **Modeling/simulation:** In machine learning or statistical analysis, a predictive model can be developed by using historical data and used for modeling a future behavior. During visualization, a user might want to simulate outcomes for a set of inputs (decisions), and might also analyze the results for sensitivity to changes in the inputs. The results can be displayed by using geo/semantic mapping that is described earlier.

Figure 1 shows the key components of the AAP, which use the Discovery (1), Integration (2), and Visualization (3) patterns.
Use cases

Now examine a use case to see how the data lake pattern is used. The example use case is chosen from the media and entertainment industry, but is applicable to various other industries.

Understanding who views one’s programs, who to target, and the response to a program is essential to any media company. Depending on the completeness of the solution, a producer who wants to increase the viewership of a program can execute multiple use cases that use the data lake search pattern.

- Multi-platform viewing: Measures a TV show’s performance across linear and non-linear viewers. Linear viewing refers to the viewing of programs that follow a broadcast schedule. Non-linear refers to the viewing of programs that were previously recorded, available on demand or from streaming services (OTT).
- Audience composition and indexing: Identifies specific audience/customer segments or audience attributes across different TV shows.
- Audience engagement and targeting: Identifies existing fans, targets new prospects, and analyzes overall engagement across TV shows.
- Social sentiment analysis: Analyzes social conversations to understand audience sentiment (positive/negative) across different TV shows.
- Social sentiment trending and correlation: Analyzes and correlates social trends to TV show performance (ratings).
- Social media indexing and visualization: Enables the exploration of extracted social media postings about TV shows through the ability to do semantic-based indexing of social media data (unstructured), faceted search of social data.

A closer look at the multi-platform viewing sub use case follows.
Multi-platform viewing use case

With the digitization of content available on the internet, TV programs no longer must be watched when they are broadcast. A TV program can be recorded to watch later (time shifted), viewed on demand from a cable or satellite provider, or streamed from an over-the-top provider. For a TV program producer, it is no longer enough to monitor the linear ratings of the TV show. It is important to also monitor the show's performance (size of viewership) across linear and non-linear channels. Linear viewing ratings are usually provided by panel-driven surveys or other sources such as Nielsen while non-linear data comes from various sources. These different data sources must be integrated and the results visualized so that the producer has a better understanding of what was watched across different channels.

Consider a high-level flow for this use case, as illustrated in Figure 2:

1. Linear and non-linear viewing data is collected. Discovery tools are applied to this data to identify the meaningful data.
2. The different data sources might need to be correlated with other data so that the Linear and non-linear viewing data is merged into a coherent view. The result of such integration can include 360 degree customer view.
3. Different visualization techniques are applied to the preceding results to enable the producer to understand who is watching the shows and what channels are being used. For the user interface to be more easily accessible from a wide range of devices, it makes sense to implement a web-based HTML5 interface with JavaScript visualization components. Targeting such a platform also ensures a high level of compatibility with many other products and tools that might already be used within an organization.

Figure 2. Flow of the multi-platform viewing use case
Let's look at how the pattern is applied to the multi-platform viewing use case:

**Discovery**

- **Business understanding:** Audience measurement determines the audience size and composition. As the first step, any project for audience measurement needs to be aligned to business outcomes. Each business outcome needs to be defined with a specific measurement. For example, a 10% increase in a specific segment of viewership results in a 20% increase in customer engagement.

  The initial step in the pattern of building an integrated audience measurement platform is to determine the business requirements of the project. To create an integrated view of how the audience responds to different content, you need to:
  - Integrate silos of data into one single repository
  - Create a detailed and deep view of each individual audience member
  - Develop ad hoc business analysis reports and to help the process of decision making.

- **Data modeling:** Data modeling usually includes the creation of three different models: The conceptual model, the logical model, and the physical model. For the modeling of a problem that involves multiple technologies, start this process with the creation of an ontology. The ontology contains definitions that are associated with the domain of audience modeling and the relationships between these elements.

  The ontology is application independent. You can use the ontology for multiple tasks. In this problem, the ontology contains the definitions for the audience, the viewership, the elements that form that audience, the segments, the programs, broadcasts, transmissions, and other concepts that are related to the definition of the audience.

  **Figure 3** shows a basic ontology definition for the audience insight platform, relating some typical concepts in the audience insight platform. In this sample ontology, we describe the audience as an element that can be composed by persons and by families. We identify that each person is part of a group that is called a family. We associated the concept of a set top box. This device has a relationship with the family, as it is normally watched by some member of the family. Normally we cannot know directly which member of the family is watching each specific broadcast. Other elements consist of the broadcast, which can be a single event, or an episode of a series.
**Figure 3. Basic ontology definition for the audience insight platform**

The ontology is used as a basis for the creation of the conceptual, logical, and physical data. The advantage to start with the definition of an ontology is that you can use the ontology as input to other components of the platform after this step. More specifically, information retrieval tasks, data mining, and discovery tasks can use these definitions to collect information and create relationships automatically from raw data.

- **Data inventory:** The data that is related to this problem is at the core of the integrated platform for audience measurement. For many years, companies like Nielsen in the US provided statistics of audience measurement and composition. Today, broadcasters have access to other sources of more data that needs to be taken into account, measured, integrated, and managed.

  The creation of an ontology is needed to document the semantic of the data objects selected. The ontology contains definitions of objects of interest in the platform. The most basic individual of the ontology is an audience member. This audience member has several attributes, which help you aggregate them into audiences. Another example of individual would be an episode. Episodes are single broadcasts, which can be part of a series, or can be individual events.

  The initial step in the creation of data audience is to review all the possible sources of information, and catalog them. This catalog consists of a repository of all the possible containers of data, and which individuals are contained in each one of them.

- **Data selection:** You determine the multiple heterogeneous data sources that can contain the data that you must extract to create a complete picture of the entity or event that is being analyzed.

  The following are typical sources of data, which can be mapped to the individuals described in the ontology:
  - Social media data: Audience member, episode, series.
  - Set top box log data: Subscriber, audience member, episode, series.
  - External ratings: Episode, series.
  - CRM: Subscriber.
  - Third-party CRM: Audience member.
Integration

• **Data unification:** This step combines all the sources of data to match the conceptual model that is created for the business problem. You need to create access to all data sources, by using federation, creating routes for the data, or integrating them. It is also important to consider the flow of information in this step.

Most big data projects rely on accessing directly the data on its source, instead of loading and transforming it. For the audience measurement platform, some sources of data might require specific software platforms to be integrated.

Two examples of this situation are social media data and set top box data. Social media data is big data. It consists of billions of individual text interactions, with approximately 500 million new inputs to be analyzed each day. For this size of requirement, you can use BigInsights, which provides a clustered environment for text analytics by using Hadoop technologies.

Set top box data is information that flows dynamically. This type of data is streaming data, and over time can accumulate in huge quantities. The approach with streaming data is to process it as it flows in real time by extracting, analyzing the relevant information, and storing the analyzed results. For this type of streaming data, you might use InfoSphere Streams.

• **Data aggregation:** When the data is organized and integrated in the audience measurement platform, the next step is to aggregate it at a logical level, matching ontology individuals with physical repositories of information.

Creating data metrics for integration is a complex problem. The main challenge of this step is that entities have different levels of aggregation across the data. Another challenge is that entities need to match across repositories.

For example, consider the problem of defining the audience for an episode. The audience for each episode consists of measurement of size and composition. The size is determined by aggregating multiple sources of data. Nielsen and other research agencies provide an estimated number of the audience by each segment. One needs to attach to each segment an estimated number of viewers from other sources of data. Consider the time shifted case: You can obtain this number from the logs and reports that are created by the set top boxes. In this case, the level of aggregation is at the individual level. However, it is common for media companies to allow multiple methods and platforms to record their programs for later viewing. In this case, you need to introduce parameters to account for the data you have from the time shifted audience. It might happen that you know for multiple studies that one platform (the one that you have detail access to) has 50% of the time shifted market. In that case, to obtain an estimation of the global time shifted audience for an episode, you need to multiply the results by two. Additionally, you need to note that the certainty of this number decreases, as you introduced an external parameter.

In other cases, one might need to make estimations not of the size, but of the composition. For example, the social media data contains multiple dimensions not available in other sources, but the data is sparse. It might lack demographic parameters like age, as those parameters are extracted over time from the data. For this reason, one might need to estimate the composition of the audience based on additional information or other sources of data.

• **Data transformation:** After you estimate parameters to transform different sources of data, you need to create a layer of calculated aggregated information. This layer includes all the
data, the coefficients for transformation, which are stored in a repository of results to be input for visualization.

With the creation of this additional layer of data, you have more flexibility at design time. For example, the layer gives you the flexibility to change visualizations without needing to know details of the layer of implementation and calculation of the intermediate results of the reports.

- **Entity analytics and resolution:** Television audience data can carry some inherent ambiguities. While set-top box data is associated with a television set in a room, often it is hard to ascertain who was watching the television. The viewer might be an adult male or female, a child, or the family dog who is watching a television that was left on. However, a non-linear viewing on a mobile device can be precisely connected with an individual. In addition, robot televisions can test the programming, and might need to be removed from the audience data. With entity analytics, you can assign probability for these diverse sets of views and identities by using the available data. For example, the programming can be used to guestimate the gender and age of the viewer. Also, the non-linear viewing can be used to establish historical viewing patterns and applied to related set-top boxes.

- **Data mining:** One of the last steps in this pattern is the data mining step. The audience measurement platform provides a single repository and view of different audience segments for analysis. Multiple variables are attached to each of the individual entities of the ontology. Multiple tasks that you can include in this step are:
  - Clustering: Grouping similar programs, similar audiences.
  - Classification: Attaching new labels to different series based on the type of audience that is followed. This task includes classifying audience members based on their engagement or social influence.
  - Anomaly detection: Identifying unusual drops or increases of viewership in specific segments of the population.
  - Regression: Modeling different episodes to forecast future viewership.

**Visualization**

- **Ad hoc queries:** A key function for interactive data analysis and visualization applications is the ability to filter the associated data so that relationship information can be more easily inferred. Accomplish this task through interactive visual components, a separate filter component, and external updates to data. Each method comes with benefits and drawbacks, and in a real-world scenario your application likely implements a combination of these methods so the user can better understand what the data offers.

- **Geo/semantic mapping:** Data visualization through graphic visualization is a key ingredient to analyzing and eventually understanding the information that you have available to your organization. As mentioned in Discovery and Integration, big data projects are likely to pull together information from a wide variety of sources. It can be difficult to present this data in a way that makes sense to people from varied backgrounds and roles within your organization without doing so with accessible and intuitive visual components. Creating these visualization components by using tools that follow HTML5 web standards is a good way to ensure a high level of accessibility from a wide range of devices. A number of excellent visualization libraries are available to make this task possible.

- **Integration:** One important aspect to visualizing data is in understanding the integration steps taken so that you can help focus the attention of your audience on these aspects of the data.
The focus of the integration process is to pull together data that is related to various entities (audience, shows, box office openings, locations, and others) and combine those entities into a coherent data model. Multiple entities have relationships that need to be easily understood when you view the visualizations that are created for your project. Another key aspect is related to the integration between the various applications, tools, and libraries that are used to create and run your visualizations. In some cases, this integration is simple — for instance, loading a single visualization JavaScript library into a stand-alone web page. Other cases might be more complex, where you must combine multiple backend data sources, multiple visualization tools, and ensure that each of the aspects work together within an existing application that comes with its own set of libraries and tools.

- **Visualization:** There are many ways to visualize data. These methods range from bar and pie charts most people are familiar with to Gantt charts, choropleth maps, box plots, chord diagrams, and many others. Deciding which visualizations to use is something that happens after you have an understanding of the type of data you are visualizing. Another consideration is in choosing the tool, library, or application that you use for generating visualizations. It is important to take into account the strengths and realities of the data you have available plus the capabilities and limitations of the tools at your disposal.

**Design considerations**

Several design points in the data lake exploration pattern can be challenging to an implementer of the pattern. We focus on design points for each category in the pattern, which you often run into when you implement the pattern. These points are not the only design areas to consider but are described here as illustrations.

**Ontologies**

Often, we classify unstructured data into categories to better understand the relative distribution across a known classification scheme. Let's use an example from online purchasing. The classic product categories originated from the Yellow Pages. However, categories are typically tree-like structures. Each node is a subclass of the node above it and can be further subclassified into further specialized nodes. For example, a scooter is a subclass of two-wheeler, while an electric scooter is a subclass of scooter. A node can be a subclass of more than one entity. A subclass shares the attributes of its super-class. Therefore, both scooters and electric scooters have two wheels. While the classic product catalogs were static and managed by administrators without organized feedback, the unstructured analytics can form a dynamic hierarchy, which you can adjust based on usage and search criteria.

A more general representation of conceptual entities is found in ontology, which is an abstract view of the world for some purpose (see Resources). Ontology defines the terms that are used to describe and represent an area of knowledge. Ontologies are used by people, databases, and applications that need to share domain information (a domain is a specific subject area or area of knowledge, like medicine, tool manufacturing, real estate, automobile repair, or financial management) and can include classifications, relationships, and properties (see Resources). With formal ontology, we can create a "Semantic Web," which can provide structural extracts to machines, thereby providing them with ability to extract, analyze, and manipulate the data.
Entity analytics and resolution

The task of the creation of audience profiles is a new task that emerged from the availability of data around audiences in social media. Before social media, the only way to obtain data about viewing habits was to use classical statistical methods to sample the population. For example, interviews, focal groups, and statistical estimation techniques.

People in social media reports, intentionally or not, frequently provide information about their viewing habits, preferences, opinions on the latest episode of their series, and what they recommend that other people watch.

In this context, entity analytics is necessary when we analyze any audience. The task of entity analytics is to create an individual profile of a person, to accumulate incrementally clues, and to resolve possible conflicts between conflicting information.

Entity resolution is the task of:

- Mapping and aggregating profiles that belong to the same person across the system
- Removing duplicate profiles for the same entity
- Accumulating context that can help resolve this situation in the future if it's not solvable today

The steps that are required to complete this task are listed next, with an emphasis of creating a pattern that can be iteratively applied when new data sources are discovered. An example also illustrates the steps.

Create an entity model that is linked to the ontology

The creation of a conceptual entity model is a step of an architectural design in the platform. When no entity model is available, there is an immediate need to create an entity model of the audience. We'll assume the latter case.

With an entity model of the audience, you can focus on the relevant information that you need to respond adequately to a business requirement.

The entity model is used to create a clear association between the data available and the business requirements. It is a superset of the original entity relationship (ER) model in the sense that this entity model additionally has different metrics that are associated to their definition. For example, level of aggregation, level of certainty, and a model to do real-time accumulation of relevant information around these entities is included in the entity model.

The difference between a conceptual entity model with a typical ER is that an ER model is focused on defining a structure for a database in a formal language. Our conceptual entity model is closely related to the ontology in the sense that it defines concepts, attaches them to content or structure of a subject matter, in a formal language, but it's not bounded to be the definition of a structure for a database. Instead, it's a conceptual model to which we will try to link different data repositories.
Associate sources of data with the entity model

The process starts from the business requirements. In this case, the main business objective is to model the audience for a broadcast scenario.

For this audience, the conceptual model has a primary entity, which would be a person, who belongs to an audience.

One of the possible sources of data to create an initial set of individuals is social media data. There are 500 million new Tweets every day.

The next step is to create social profiles that are based on this input.

Figure 4. Social profiles in InfoSphere BigInsights

BigInsights contains a set of tools that allows a user to create simple social media profiles by using Twitter data, which is based on the content of the Tweets. Using this tool, or a similar tool, you create a baseline of profiles for your target audience.

The profiles are created from Tweets from multiple topics. The result is initially not going to be skewed to collect profiles relevant to broadcasting or television. To analyze only those profiles of interest to you, you need to analyze the conversations and perform a topic classification on the subject you want to analyze. With BigInsights, this step is simplified, as the analysis tool has a built-in set of seed words to analyze media and entertainment companies, including names of companies, series, television shows, and other information. Using the specific module from SDA
can accelerate your development time, removing the task of creating a specialized set of keywords to select and apply them to the social profiles created.

**Associate metrics to the data by using the entity model (ontology)**

One key step in the design of the audience metrics is to clearly identify each data source entity, and link them to the entity model. Assess each of the following metrics per each entity, in an iterative process that links a business requirement and the final data source:

- Level of aggregation
- Level of certainty
- Level of time aggregation
- Governance policies of the data
- Confidentiality

**Entity resolution**

When each of all the data sources is identified, linked to the entity model, and each attribute is measured against the defined metrics, you need to remove duplicate entities.

**Figure 5. Entity Resolver in InfoSphere BigInsights**

Entity resolution is the process of combining multiple data sources to disambiguate real world entities from different sources to one single conceptual entity. This step is complex, and requires the use of both deterministic and probabilistic methods to achieve acceptable levels of confidence.

There are multiple ways to perform this step, and each method is designed to consider the characteristics of the data that is being handled and disambiguated.
The task of entity resolution started decades ago as the task to disambiguate multiple records in a data storage repository. Multiple algorithms perform this task. Most of them can be classified in the following groups: deterministic and probabilistic.

- Deterministic methods rely on algorithms that are created specifically to match entities, considering the special characteristics of the data sets being analyzed.
- Probabilistic methods rely on statistical methods to determine the likelihood of a linkage between a set of entries in the real world to one single conceptual entity.

In the context of audience insight, the following problems need to be addressed by entity resolution:

- Matching of CRM records
- Matching of social media records
- Matching of CRM records with social media records

**Integrating multiple visualization tools**

You need to consider various areas when you design the visualization. Many articles on developerWorks and other publications describe areas to consider as you create effective visualization interfaces and you’ll want to review these articles. An area that is not addressed well is the integration of alternative visualization tools, which is often necessary. The section discusses design considerations for integrating two visualization tools and the provided example uses Watson™ Explorer and D3.

The likelihood is high that your visualizations need to integrate with an existing application, whether it is to use some data source, be shown on an existing website, use (or at least co-exist with) a specific version of some library, or another scenario. In our example, we discuss visualizations that are created with D3 and configured to run within the Watson Explorer pages. At the same time, we combine these visualizations with others created with the Highcharts JavaScript library to use the tightly integrated features of this library within the Watson Explorer pages. By taking this approach, we can continue to use some of the interesting visual and data manipulation capabilities of D3 while still benefiting from the built-in visualization capabilities of Watson Explorer.

Watson Explorer makes the integration between the YAML-based widget and the Highcharts API seamless by providing widgets that are dedicated to these types of visualizations. But when you use a different third-party library such as D3, you need to use one of the other two widgets that are provided for this type of task: the Iframe or iWidget widgets. You can use the Iframe widget, a basic widget that provides a view into another web page that contains the D3 elements. This integration will, for most purposes, be transparent to the user.

A wide selection of visualization tools are available. Since Highcharts already has excellent integration with Watson Explorer, new projects on this platform benefit greatly from targeting this library for their visualization needs. When you use the pre-existing visualizations that are built with the D3 library, you can also use the data manipulation aspects of D3 to get the data into JSON objects. The objects are easy to plug into your charts and graphs. You can choose to reuse both the visualization and data manipulation aspects of D3 within Watson Explorer as outlined earlier.
If you add D3 components to a Watson Explorer page by using the Iframe widget, it is important to understand limitations to how your visualization components can interact, both with each other and with any other HTML element on the parent page. Charts behave as if they are on separate pages, so components that share filters or that react to changes on another portion of the page do not work without connecting each component in some way. One simple way to create this connection is through Iframe attributes whose values can be set programmatically by a script that runs on the parent page. Another more functional way is to connect each visualization to a shared remote data source through data binding. The data source tracks selected data filter options from all components, and any updates are automatically applied to any other widget that uses the same data source. There are several ways to do data binding if you are implementing it from scratch. Some services provide a library and hosting to make it simple (Firebase, MongoDB, and others).

**Tools and technologies**

This section provides a brief description of the tools that are described in this tutorial that were not covered in previous tutorials.

**Watson Explorer**

Watson Explorer provides search, navigation, and discovery over a broad range of data sources and applications both inside and outside your enterprise to help you uncover the most relevant information and insights.

Watson Explorer delivers best-in-class search features to enable users to access what they're looking for quickly and efficiently. Features such as synonym support, spotlighting, relevance tuning, custom dictionaries, robust syntax and "did you mean?" enable productive searching without the frustration that often results from less capable search.

**Information Server**

Information Server creates and maintains trusted information to support strategic business initiatives that include big data, point-of-impact analytics, business intelligences, data warehousing, master data management, and application consolidation and migration.

- Enables collaboration to bridge the gap between business and IT
- Helps align business and IT objectives
- Provides metadata integration & data lineage insight
- Always-on operational data integration & data quality
- Linear scalability and infrastructure optimization
- Broad connectivity to nearly all data sources
- Productivity tools for organizational efficiency
- Accelerated data integration deployments

**BigInsights**

IBM® InfoSphere® BigInsights™ brings the power of Hadoop to the enterprise. Apache Hadoop is the open source software framework, used to reliably manage large volumes of structured and unstructured data.
IBM makes it simpler to use Hadoop to get value out of big data and build big data applications. It enhances open source technology to withstand the demands of your enterprise, adding administrative, discovery, development, provisioning, security, and support, along with best-in-class analytical capabilities. The result is a friendlier solution for complex, large-scale projects.

InfoSphere BigInsights, for Hadoop, empowers enterprises of all sizes to cost-effectively manage and analyze big data — the massive volume, variety, and velocity of data that consumers and businesses create every day. InfoSphere BigInsights helps increase operational efficiency by modernizing your data warehouse environment as a query-able archive. You can store and analyze large volumes of multi-structured data without straining the data warehouse.

**Cognos Business Intelligence**

Use Cognos® Business Intelligence to make better and smarter business decisions faster with solutions that take business intelligence (BI) to a whole new level. Innovations in BI from IBM provide broader analytic capabilities so that everyone has the relevant information they need to drive your business forward. IBM business intelligence products are designed to integrate with one another and with many third-party solutions, including leading big data platforms. You can start to address your most pressing BI needs almost immediately with the confidence that you can seamlessly grow your solution over time to meet future requirements.

**Conclusions**

This tutorial discussed an exploration pattern for data lakes that uses the Advanced Analytics Platform (AAP) and the key components of the pattern, including discovery, integration, and visualization. Through a use case from the media and entertainment industry, you learned about the implementation of this pattern and design considerations typical for the pattern. The next tutorial in this series shows how analyzing large volumes of data at high velocity, resulting in real-time actionable steps, is an essential component of any analytics platform. Dealing with such data poses some specific challenges, and the tutorial also describes how to approach these challenges.
Resources

Learn

- Learn more about ontologies:
  - Read about the integration and semantics of activity modeling in *Representation of Activity Knowledge for Project Management* by A. Sathi, M. Fox, and M. Greenberg (IEEE Transactions on Pattern Analysis and Machine Intelligence 7, no. 5, May 1985).
  - Learn about usage scenarios, goals, and requirements for web ontology languages in *OWL Web Ontology Language: Use Cases and Requirements* by Jeff Hefflin (W3C Recommendation, 10 February 2004).
  - Visually represent your data with the open source D3 JavaScript library. (Bilal Siddiqui, developerWorks, July 2013):
    - **Data visualization, Part 1: Visualize browsing metrics with SVG and D3**: Apply D3 to visualize social media data.
    - **Data visualization, Part 2: Use D3 component layouts**: Focus on the concerns that inform the arrangement and design of your visualization application while expanding on the use of D3.
  - Learn more about *Watson Explorer*. Reduce your research time and work smarter and faster.
  - Learn more about *InfoSphere Information Server*, a data integration platform with the tools so you can understand, cleanse, monitor, transform, and deliver data.
  - Learn more about *InfoSphere BigInsights*. Use BigInsights and Hadoop to manage and analyze big data.
  - Learn more about *Cognos* and creating enterprise-class reports, independently assembled dashboards, based on trusted information, mobile BI, and much more.
  - Learn more about *D3*, a JavaScript library for manipulating documents based on data.

Get products and technologies

- **Evaluate IBM products** in the way that suits you best: Download a product trial, try a product online, use a product in a cloud environment.

Discuss

- Get involved in the *developerWorks community*. Connect with other developerWorks users while you explore the developer-driven blogs, forums, groups, and wikis.
About the authors

Dr Arvind Sathi

Dr. Arvind Sathi is the World Wide Big Data Architect at IBM. Dr. Sathi received his Ph.D. in Business Administration from Carnegie Mellon University. He is a seasoned professional with more than 20 years of leadership in information management architecture and delivery. His primary focus has been in leading big data and advanced analytics architecture activities for IBM clients in telecommunications, media and entertainment, and energy and utilities organizations, worldwide. He has published three books on big data analytics.

Julio Wong

Julio is a certified IT Specialist. For the last seven years, he has worked in the Media and Entertainment Industry, at the Global Solution Center. He is involved in technical sales support activities such as building proof of concepts and customer briefings. His focus has been in Media Business Process Management solutions. In the past, he worked on other sectors such as Government, Insurance, Financial, and Industrial.

Chris E Loza

Chris Loza is an IT Specialist with experience in the development of solutions for text analytics and big data. He works as a data scientist for the Global Solution Center, finishing his PhD with a dissertation on unsupervised semantic analysis for social media. He has published papers and IBM internal publications in the area of social media analysis, on the subjects of sentiment analysis and the creation of integrated solutions to analyze customer behavior.

Mathews Thomas

Mathews Thomas is the lead architect for Communications Sector at the IBM Global Industry Solution Center (GISC) which hosts the primary IBM industry solution center and labs in North America for the telecommunications, media and entertainment, and energy and utility industries. The GISC is also an IBM Analytics Solution Center where Mathews is the lead architect.
Joshua Purcell

Joshua Purcell, an IT Specialist, focuses on data visualization and application integration. Most recently he has worked with various visualization tools and libraries to visualize and analyze data that is pulled from various IBM products such as Netezza, Watson Explorer, and others.

© Copyright IBM Corporation 2014
Trademarks
(www.ibm.com/developerworks/ibm/trademarks/)