Data Planet & Data Galaxy

A Call to Awakening –
Data, Everywhere, Everyone

WHITE PAPER | July 2024

CONTENTS

CAN WE REALLY CATCH UP WITH DATA? 2

1 The Heavy Weight of a Closed System 3

2 Blurred Data Ownership 5

3 Out of the Box: Unleash the Value From Diverse Data 6

OPEN ECOSYSTEM: A PARADIGM SHIFT 8

1 Safeguarding Data Ownership 9

2 Live & Ubiquitous Capacity 10

3 Data for Everyone 12

4 Data Synergy 13

SUMMARY 15

Embracing a Data-Centric Future with a New Open Data Ecosystem **15**

Can We Really Catch Up With Data?

By now, professionals of almost every field would agree the era of the Data Economy is upon us. After decades of computerization, buttressed by the burgeoning Internet of Things (IoT) and 5G networks and mobile devices, tiny and numerous sensors are now densely distributed throughout the earth. Data on human activities and the environment are constantly generated and circulated by these mobile devices. Behind this remarkable progress is an unprecedented group of forces that echo each other and converge: high-speed cloud computing capabilities, vast amounts of data, and thriving artificial intelligence (AI) applications have all fed into one another's acceleration in the past two decades. These forces continue to inspire mankind to innovate for the future, making the overall economy's appetite for data ever more ferocious.

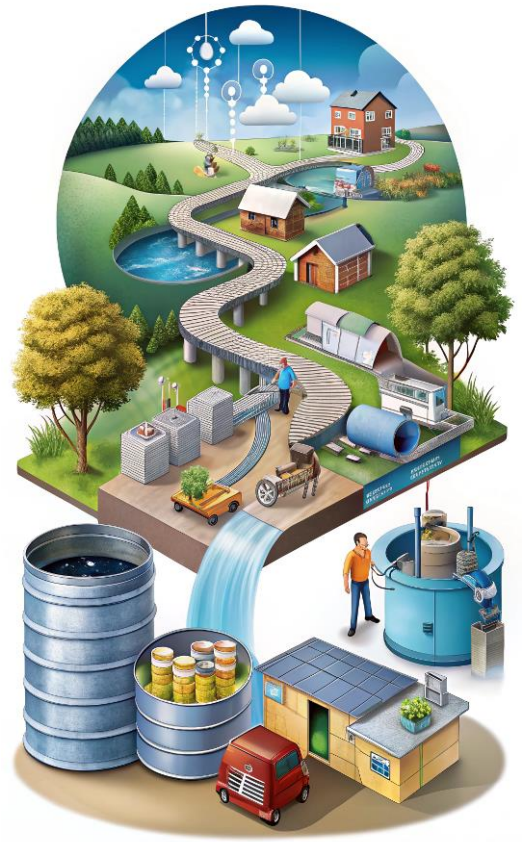
Data exist as an asset in today's society. Some liken it to the new oil. However, while many people benefit from the new business opportunities brought by data, only few have the technical expertise and privilege to take full advantage of it. Most people still follow the existing ways of working and learning, unaware of the value embedded in data. In today's world of ubiquitous data, we believe if everyone had the ability to understand and utilize data, it could become an indispensable source of competitive advantage in business and upward mobility for the economy as a whole. How to reach the full potential of this new asset for everyone, therefore, becomes imminently essential.

To understand the potential issues behind the current use and provision of data, we compared the differences between current data platforms and what a true data ecosystem should entail: a balanced data system where demand and supply interact smoothly, efficiently achieving a state of equilibrium under market mechanisms. The resulting discovery is a surprising contradiction between today's open-ended nature of ubiquitous data and the prevailing closed-ended treatment that they receive. Due to the protection of personal information and data assets, fixed management orientation and technical thresholds, most data owners lack the incentives to share data. On the other hand, data consumers are trapped in the cumbersome inertia of collecting, cleaning, and integrating data. The time-consuming and laborious process of a closed system deeply hinders the possibility of widespread use of data.

1 The Heavy Weight of a Closed System

To understand the closed data processing practices that we are familiar with today, we need to trace back to the dawn of computer applications in the early 1990s. That was before the internet age, when the type of data stored in computers were mostly transactional between vendors and customers, financial records, or operational industrial data. They were organized in formats with clear column and table structures and stored in files or databases. People followed the compute-centric mode to gather and prepare data before performing the analytical work. This structure meant that the data user must first download the relevant data, which faced growing difficulties as data privacy acts became roadblocks to such practice.

Then came the age of IoT where analogue-to-digital conversion and image monitors became increasingly common. Characteristics of data - live, ubiquitous, and diverse - have also presented a new phase. Be it cars, watches, or home appliances, more objects are now connected to each other via networks. New data are sent and stored constantly. However, the compute-centric approach from the past has presented our data environments with many data silos, which greatly limit the resources and efficiency for analytical work. Today's open environment of fast-and-furious data generation greatly contrasts the closed-ended systems that want to analyze them. The processing models that were once useful in the past now seem clunky and slow.

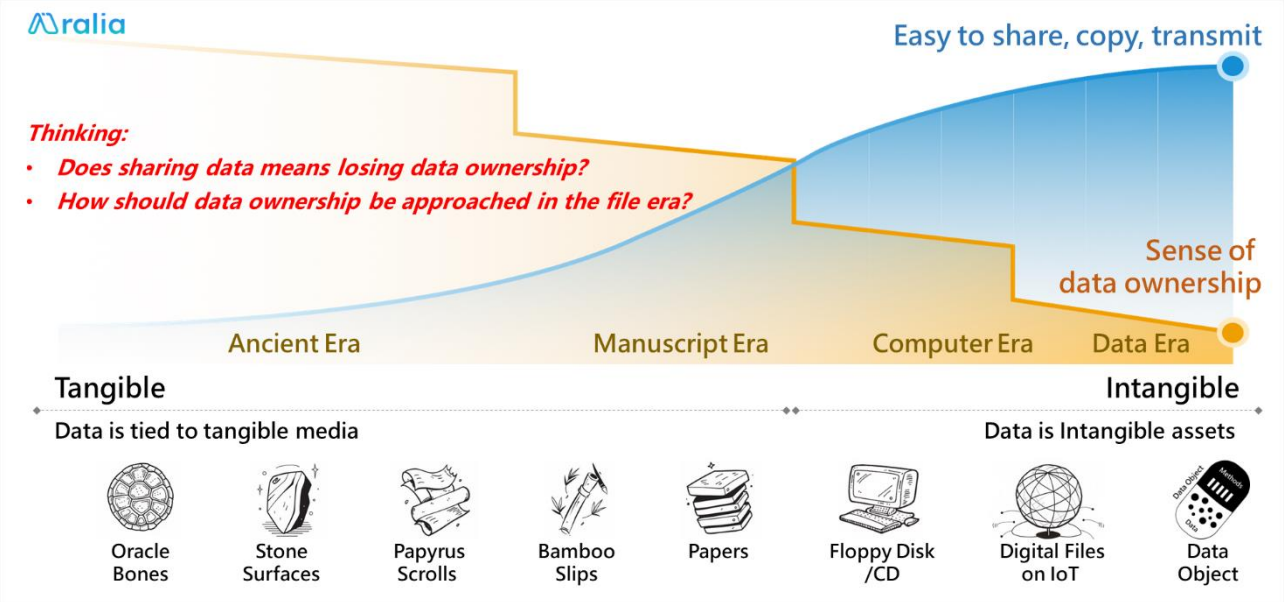


Let's imagine for a moment that closed data processing is like water usage in rural villages. Everyone goes to a nearby water source like a river or spring to fetch water, carry it back, and treat it before use, over and over again. This is repeated by every household. An open data system is like urban infrastructure; once built, water comes to every household via a faucet whenever the water is needed. This not only saves time, but also reduces repetitive work.

Moreover, in recent years, people often refer to the "shared economy" or "platform-economy" as the force that has subverted many aspects of our lives with new business models. While websites for shared homes, shared rides, social networks, and e-commerce provide unprecedented convenience services to us, they have also quietly collected consumer data such as preference and pricing acceptance levels to refine their sales strategy with precision. These technology giants, representing almost a quarter of the entire stock market valuation, now dictate the market with those data, all in the name of a shared economy. But behind these business models is the reliance on closed data models, which in turn leads to data fragmentation and monopolization.

2 Blurred Data Ownership

The history of human data has gone through the ancient era, the manuscript era, the computer era, and the current data era. As shown in Figure 1, during the transition from recording media on ancient murals to paper transcriptions, the advent of printing technology greatly accelerated economic growth. It gradually transitioned from being tangible to today's intangible digital content. Various magnetic disks, DVDs, flash memory drives, and networks have made communication faster and cheaper. Business models were made agile and vibrant. Unfortunately, these advancements in content creation, replication, modification, and easy transmission and sharing have all led to the blurred definition of data ownership.



Ownership is an integral aspect of human nature: to possess, control, and trade valuable resources. It forms the foundation of an economic framework of social order as guided by a legal system. It reduces conflict and social costs and promotes cooperative coexistence for collective growth and prosperity. However, for an intangible asset such as data it is difficult for people to clearly claim ownership. Data owners think they possess ownership of data, but in fact they are powerless when it comes to protecting their data assets. They can only protect their rights and interests by limiting their data sharing.

Data quality and usability is highly correlated with authenticity, which is inextricable from its origins. Data comprise recorded pieces of information that can be mixed and matched together into one story. While every data user expects authentic data, it is hard to decipher which data are whose. Once exchanged, data are easily taken out of context, copied, or even reshaped. After being converted and transmitted many times, the authenticity of the data becomes increasingly unclear, the source becomes difficult to trace and its usability becomes doubtful.

3 Out of the Box: Unleash the Value From Diverse Data

Closed systems and unclear data ownership limit the scope and potential of data sharing. If these restrictions can be removed under the premise of allowing data to openly interact with each other, without physical copying and moving, the result will likely be unprecedented and phenomenal. New synergies are like creating opportunities between people through dialogue or stimulating ideas across cultures. At present, academic institutions, corporate organizations, and government agencies all have vast and multi-faceted data. Unfortunately, they are isolated from each other, revealing limitations and the inability to unleash the infinite potential within.

To break free from these limitations, we must realize that past perceptions and habits around data must evolve from computer-centric to data-centric thinking. Creating an open data ecosystem is essential - data providers must be protected so they lower their guard against sharing data, and

potentially even become motivated to actively improve data quality for rewarded usage growth. Then, data users will no longer be boxed in by limited access to data. More diverse and heterogeneous ready-to-use data can be obtained at any time, like goldmines for them to explore. A healthy supply and demand will eventually lead to a balanced and sustainable data ecosystem.

An open data ecosystem forms the infrastructure of the data industry, which in recent years has grown at the same time as generative AI, but on separate tracks. Data in Latin, Datum, means recorded facts. Whether it is human intelligence or artificial intelligence, if there are no data there are no facts, and no knowledge can be learned to form intelligence. Generative AI relies on training from extremely large amounts of data and is seemingly omniscient, but it can only make predictions corresponding to the data on which it is trained. It cannot know what is happening around the world, bit by bit, on its own. Data are the raw material of artificial intelligence. Going forward, generative AI and an open data ecosystem shall be interdependent and complementary. This combination shall lead us to a healthy data economy from which all stakeholders would benefit.

Open Ecosystem: A Paradigm Shift

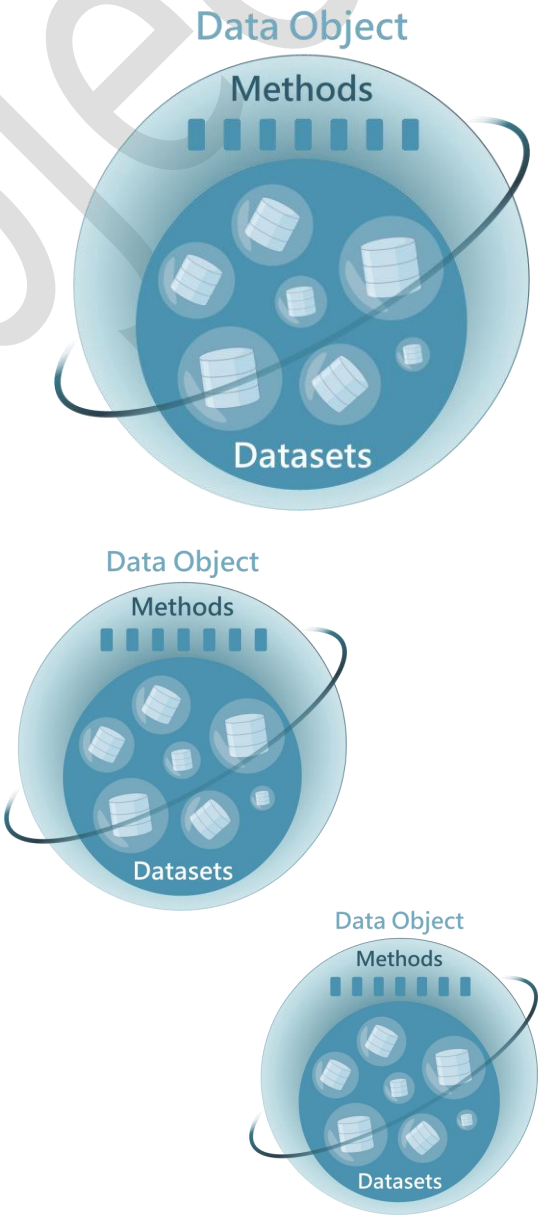
The aforementioned computer-centric mindset resulted in blurry data ownership and a closed-ended data handling strategy. By comparison, a data-centric mode facilitates a completely different, open-ended data ecosystem that encompasses heterogeneous and diverse data sourced with clear ownership from all directions, and which are constantly streaming and refreshing. Data providers and users participate, interact, and transact in the ecosystem at will. The ebb and flow of activities and gaining or losing on the values of data less or more eventually reach an equilibrium in such an environment. All data entities are assets capable of operating independently. With the authorization of data owners, users can go from operating within a single data entity to immersing in an analytical experience and discovering synergetic insight across interoperable data entities.

To ensure the success of implementing a next-generation open data ecosystem under Aralia, we necessitate a combination of four pillar principles: Safeguarding Data Ownership, Live and Ubiquitous Capacity, Data for Everyone, and Data Synergy. Each addresses a key aspect of the holistic data ecosystem. We believe that the establishment of this framework can incite a shift in the data application paradigm and usher in the long-awaited era of a shared data economy.

1 Safeguarding Data Ownership

Compared to tangible assets, intangible ones are far more easily cloned. In the early stage of the Internet, when regulation and protection mechanisms were scarce, digitized music and the film industry were profoundly impacted. Rampant infringement violations brought an array of legal instruments, including cyberspace intellectual property (IP) laws, copyrights, and trademarks. However, unlike the clear definitions of patent, copyright, and trademark, data comprise bits and pieces of information that make ownership difficult to discern. Protecting data ownership from unjust duplication, possession, tampering, or other abuses has always been a challenge.

In the Aralia framework, data are rendered into an object in the same sense of the object-oriented paradigm where datasets are encapsulated within a set of defined interfaces, called a Data Planet. Consumers conduct explorative analysis by interacting with data through these sanctioned interfaces only. This way, data owners are ensured the integrity of their data and the control of ownership without risking any tampering with are abuse of their data. This creates an incentive for data owners to come forward and take part in unleashing the invaluable benefits of the shared data economy.



As the primary principle of Aralia, data ownership fosters trust between data producers and consumers equitably via secure data sharing and without compromising the underlying data. It enables the parties of interest to proactively participate with proprietary data while reaping the synergetic benefits of sharing one another's data. Data are no longer subordinate to computing resources; they become one informational data object as an asset.

2 Live & Ubiquitous Capacity

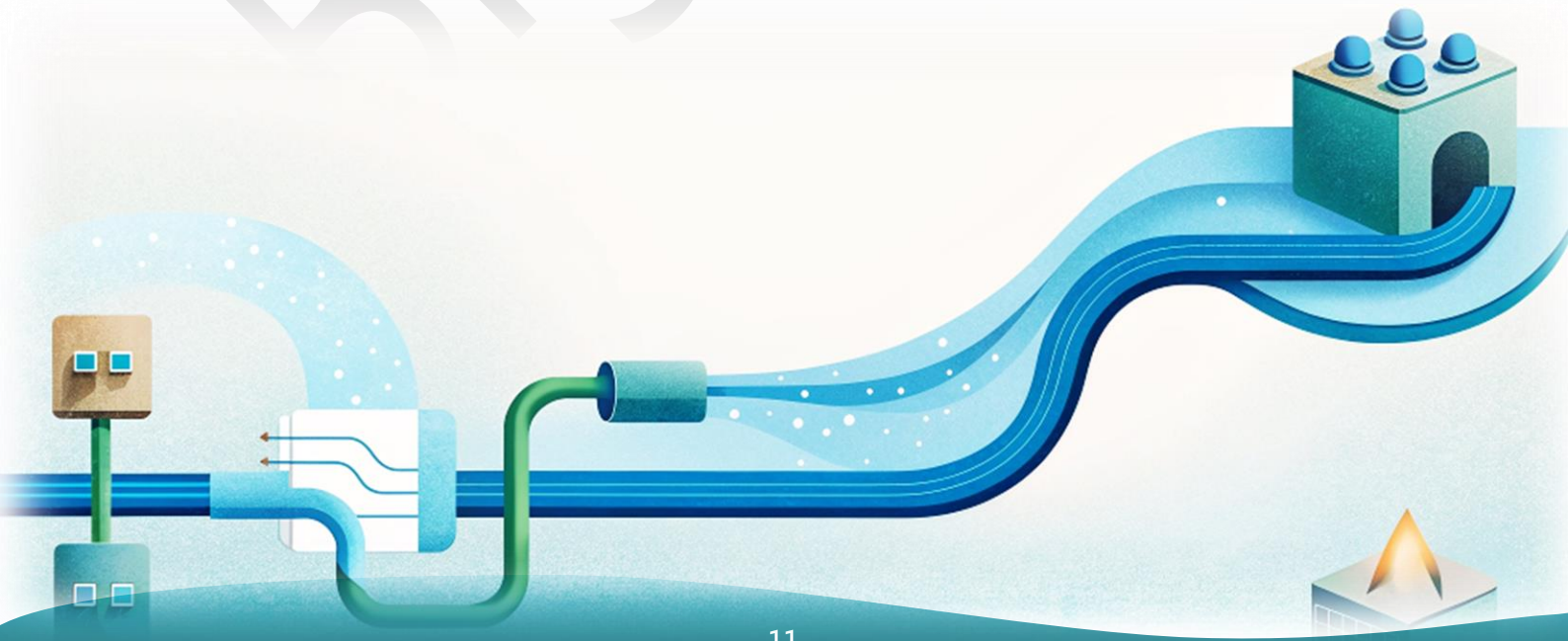
The second principle of Aralia focuses on the ability to effectively and readily provision usable data. In Aralia's open data ecosystem, the Data Planet is the most basic unit that represents the territory of every data owner's sovereignty, representing the emerging nature of live and ubiquitous data for its readiness. Although it follows the closed-ended operating mode of a series of data preparation tasks, starting from collection, cleaning, integration, to storing data, it manages to achieve the open-ended state of Ready-to-Use continuously. With a high level of usability, Data Planet's users can enjoy the most up-to-date data at all times.

Maintaining an environment of high-quality data not only benefits users, but also motivates data providers to keep their data in a ready-to-use mode. Conversely, if a Data Planet is unable to maintain such quality, it will result in its loss of competitiveness, and eventual viability. The ability to provide usable data is the key factor determining whether an ecosystem can sustain or not. Reaching the ever ready-to-use status is onerous due to the variable requirements for data quality across different domains.

Four quantitative indices measure data quality and readiness for use in Aralia: timeliness, validity, completeness, and stability. Data providers must disclose recent index readings and historical trends, if applicable, while availing their Data Planet for others to use.

Although the time-consuming and labor-intensive nature of data preparation remains the same, the fundamental difference is whether the job is done in the groundwork-based data processing mode for open-ended usage by data providers, or in project-based data processing mode for closed-ended usage by data users. All too often, closed-ended data gathered for one-off research grants or projects are shelved without a continuing life once the project concludes. The underlying support data and report dashboard almost always become obsolete. On the other hand, the open-ended groundwork-based data preparation, maintained by its owner, is done once by establishing the data pipeline for continuous use by many users, greatly improving the economies of scale for all.

With data ownership secured, the ever readiness of data becomes a motivating responsibility of owners. Therefore, Aralia data can be supplied and consumed like electricity and water at the flip of a switch.



3 Data for Everyone

The use of prevailing business intelligence (BI) tools requires prior knowledge of relevance between datasets and pre-aligning them for intended analytical purposes so decision makers can visualize results in a streamlined, customized dashboard. The impediment to this process is that it is frequently trying to keep pace with the constantly changing external environment.

Through Aralia, we endeavor to deliver an intuitive inference style of data exploration for curious minds, facilitating an immersive experience for everyone. Data consumers can drill down into any dimension and pivot against other dimensions flexibly. Through this property, Aralia lifts the limits and lowers the barrier to analysis for consumers.

Under the Aralia framework, SQL or programming skills are no longer required for analytic work. Any logical thinker can be an analyst and draw inferences through descriptive data analysis, cross-referencing, anomaly detection, and trendlines within a single or across multiple Data Planets spontaneously. As user ideas progress, if equipped with congenial data and tools, the unknown future will be full of rewards and surprises.

Aralia emphasizes that an open data ecosystem should be easy to use, so anyone can engage in it regularly at work or in life with very little barrier. We envision data democratization where data will be as accessible as computers and the Internet are today. We believe that only through opening access to data for everyone can efficiency be achieved or upgraded for every individual, enterprise, and society as a whole.

4 Data Synergy

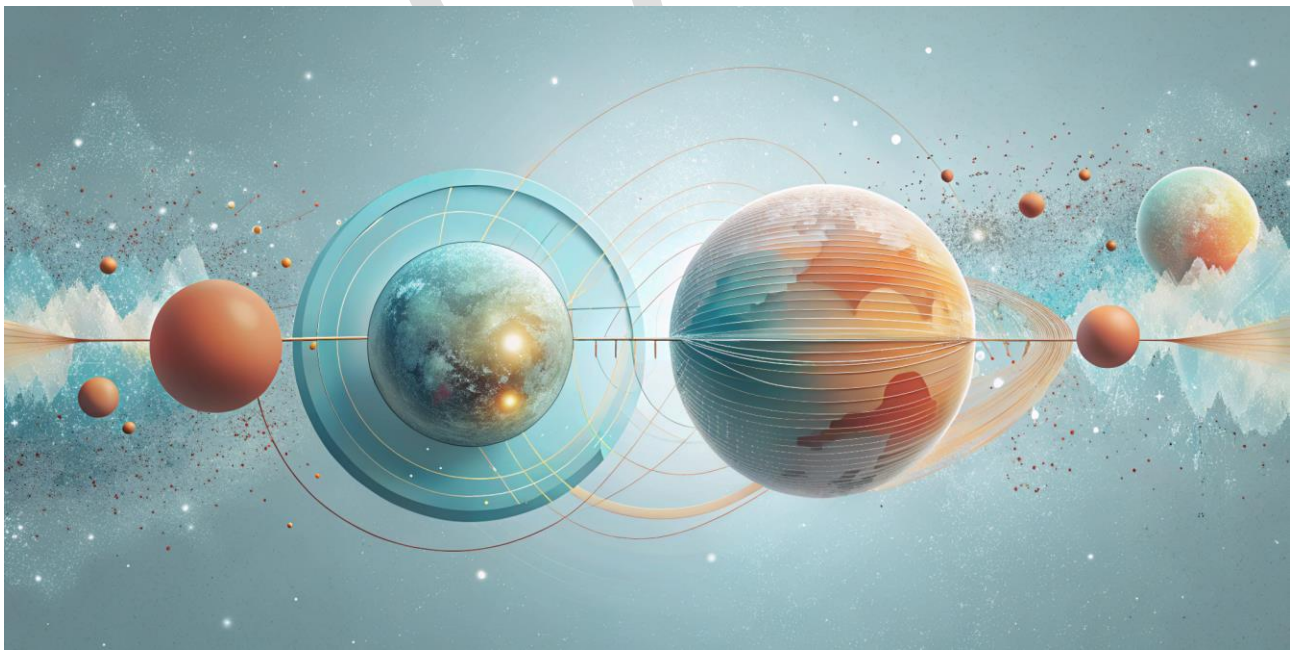
In a completely open data ecosystem, a plurality of planets with diverse and related data that uphold data ownership principles shall have opportunities to encounter one another in cross-referencing. As the number of data planets increases, opportunities for cross-referencing from existing planets to new ones would grow by an order of magnitude, and open up vast possibilities to find new synergetic insights. Like a jigsaw puzzle, it is not until an analyst puts every piece together that the whole picture is revealed. Facilitating the interaction of diverse data for synergetic insights is an important mission for an open data ecosystem.

For example, a green energy generation plant can reasonably glean insights from seasonal local weather reports, temperature, level of photons (sunshine), and wind conditions to effectively plan its production and distribution. Such a power plant may further benefit from data on atmospheric pressure, humidity, or topographic features from heterogeneous sources to further hone in on forecasting future energy generation, along with demographic data, industrial consumption, and economic trends.

With such diverse data, the project-based approach for a closed-ended user mode would require a data platform capable of processing billions of records, collecting from different data sources, and cleaning and storing data before analysis can even begin. Conversely, under an open data ecosystem, analysts need only explore, cross-reference, and derive the answer in a short amount of time. It is our belief that data users should invest most of their time on analysis rather than on costly data preparation.

To address the need for synergetic insights, the Aralia framework offers the cascading style of exploration within or across data objects, enabling analysts to hop from one data object to another seamlessly and continuously, while building on prior findings with further data. Under the Aralia framework, closed-ended platforms and data silos will have the opportunity to become integral to what used to be the "outside" world, enriching data access without the need to unify the underlying raw data.

An open data ecosystem can unveil synergetic insights that any single data source cannot. Encounters among data planets shall attract more participants into the ecosystem to share and gain the value of data, forming a benign cycle for data providers and users alike. Aralia aims to inspire the wonder of data for analysts with all the endless possibilities of discovering unknown related stories, and restoring past events with scattered evidence. The state of sustained supply and demand equilibrium of an open data ecosystem shall foster change in the way people use data, and lead us to the next threshold of civilization.



Summary

Embracing a Data-Centric Future with a New Open Data Ecosystem

The prevailing data ecosystem was established with the computer-centric mindset derived from the computerization era in the 1990s. But to keep pace with today's digital transformation and groundbreaking developments in data science, the practice of project-based data processing based on a computer-centric mindset is gradually being replaced by data-centric groundwork-based data processing. We can see that a paradigm shift is on the horizon.

It is critical that we recognize data from a fresh perspective. Data should no longer be subordinate to computing resources. Instead, it should come with a set of standardized access interfaces that can prevent gratuitous abuse and protect value for its owner. It is through the implementation of recognizing ownership that data owners are incentivized to let their data into the ecosystem. Adhering to the mindset of data centrality, data are in the hands of data owners and processed and computed using resources under the Data Planet's jurisdiction. Data users will not need additional resources to find the synergetic insights. The overall value of data increases based on positive interactions between supply and demand.

Inspired by the unlimited potential of collective intelligence behind data, Aralia hopes that adopting the four fundamental principles - safeguarding data ownership, live and ubiquitous capacity, providing data for everyone, and data synergy – will create a balanced, adaptive, and resilient data ecosystem with the potential to usher in the era of an equitable, just, and super-civilized Data Economy for everyone.

Data Planet & Data Galaxy
Change The Way People Use Data

BigObject | Aralia

Singapore

Address: 11 Biopolis Way, #09-03 Helios, SG 138667

Phone: +65 8154 4770

Taipei

Address: 4F, No. 9, Sec. 1, Zhongxiao E. Rd., Zhongzheng Dist., Taipei, TW

Phone: +886 2 2343 5562

Email: sales.aralia@bigobject.io