CLOUDERA

WHITE PAPER

# CLOUDERA DATA PLATFORM – PUBLIC CLOUD SECURITY

CDP PUBLIC CLOUD

# Table of Contents

## Cloudera Data Platform – Public Cloud Security

The Cloudera Data Platform (CDP) Public Cloud is a Platform-as-a-Service (PaaS) offering with an architecture that is based on industry standard best practices, and ensures that our customers' data and infrastructure is secure in the public cloud. CDP employs a control plane architecture which divides the infrastructure of the platform between orchestration components that reside on Cloudera's public cloud infrastructure, and data storage and processing resources within the customer's public cloud accounts. This provides you the agility and elasticity of the cloud, without the burden of learning provider-specific automation and security tools necessary to scale at enterprise levels.


Image 1: Cloudera Data Platform (CDP) Public Cloud

### Control Plane

First, let us dive into the CDP Control Plane, which is a single pane of glass, bringing together the CDP infrastructure across all of your public cloud vendor accounts (in fact, across all of your CDP deployments). The Control Plane is where the majority of management and administration interactions with CDP will take place. Provisioning of new CDP resources, such as Data Hub clusters, virtual data warehouses, and machine learning workspaces, and authorization to those resources can all be done through the Control Plane interface. Although the Control Plane runs on Cloudera's infrastructure, the power behind CDP is the seamless operation between the Control Plane and the resources that run within your public cloud account. Our PaaS offering ensures you still maintain control of your compute environments should any down time occur in the Control Plane. This means workloads are unaffected and will continue to operate within your public cloud infrastructure. With Cloudera's goal to give customers more control, CDP's public cloud architecture has been designed to ensure that customer data is not sent to Cloudera's infrastructure and remains within your workload environment at all times.

### Infrastructure

Next let's talk about the infrastructure CDP provisions inside your environment. CDP orchestrates all the necessary compute, storage, and workload infrastructure within your public cloud account. The backbone of this infrastructure is a Data Lake where all of the services for Cloudera's Shared Data Experience (SDX) run. Whereas the Control Plane manages CDP's compute and storage resources, the SDX services are how you will implement security and governance policies to control access to your organization's data. This allows policies to propagate across all workloads that you may use in CDP.

When a customer accesses CDP Public Cloud experiences, they are directly accessing the infrastructure running in their own public cloud account.

### CDP Experiences

You may hear us at Cloudera refer to the experiences in CDP Public Cloud. These Experiences describe products that reflect the five stages of the data lifecycle (i.e. DataFlow, Data Engineering, Data Warehouse, Operational Database, Machine Learning). These Experiences are Kubernetes deployments within your public cloud infrastructure that cater to a specific part of the data lifecycle. In fact, when you use these experiences, they are directly accessing the infrastructure running in your own public cloud account. This means that your sensitive data is being processed and presented to you through resources provisioned in the same public cloud account as your data.

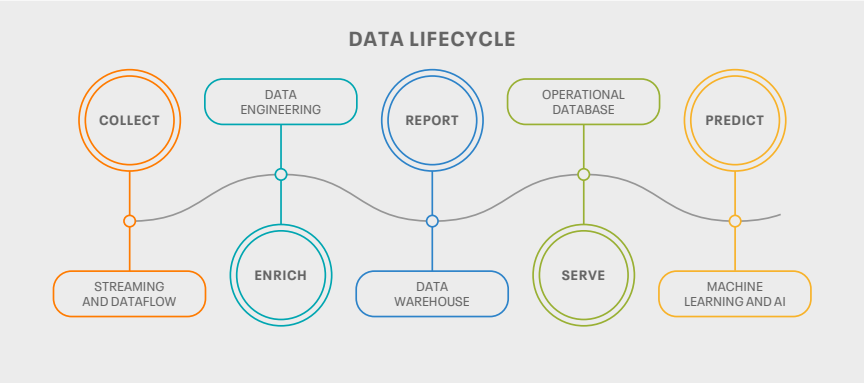



Image 2: Data Lifecycle

## Designed with Cloud Security Best Practice in Mind

The industry standard control plane design requires the infrastructure within a customer's public cloud account to reach out over the public internet to Cloudera's systems. CDP utilizes several methods to ensure these connections are secure from outside threats. Additionally, you can opt-in to extra security by limiting your infrastructure to only allow outbound connections to permitted addresses.

### Orchestration

In order to enable all of the powerful orchestration necessary for an enterprise data platform, Cloudera requires a dedicated role (depending on the cloud vendor: an AWS role, Azure identity, or GCP service account) that enables the Control Plane to perform the necessary operations within your public cloud account. Using a dedicated role is the preferred method of cloud vendors for several reasons:

- The dedicated role is mapped to one external user, which is a machine user that CDP orchestrates your public cloud account's resources with.
- The access and abilities of the cross account role can be limited in order to minimize security risks.
- This also means that the dedicated role can be disabled or removed at any time to disconnect your public cloud infrastructure from CDP.

## Monitoring

Along with orchestration, the CDP Control Plane offers comprehensive monitoring of the technologies deployed; the infrastructure that resides within the customer's account continuously reports its health status back to the Control Plane. By default, CDP provisions infrastructure to use private IP addresses. Cloudera Connectivity Manager (CCM) acts as a bridge between the infrastructure protected behind private IPs and the public endpoints of the Cloudera Control Plane by utilizing an encrypted NAT-T (Network Address Translation-Traversal) tunnel. CCM does not allow inbound requests and instead only requires outbound access to the public internet in order to function. If you choose to use public IPs for your CDP infrastructure, CDP creates public endpoints URLs and limits communication to HTTPS. Rest assured that all traffic between your public cloud account hosting the workload environment and the Cloudera Control Plane is encrypted.

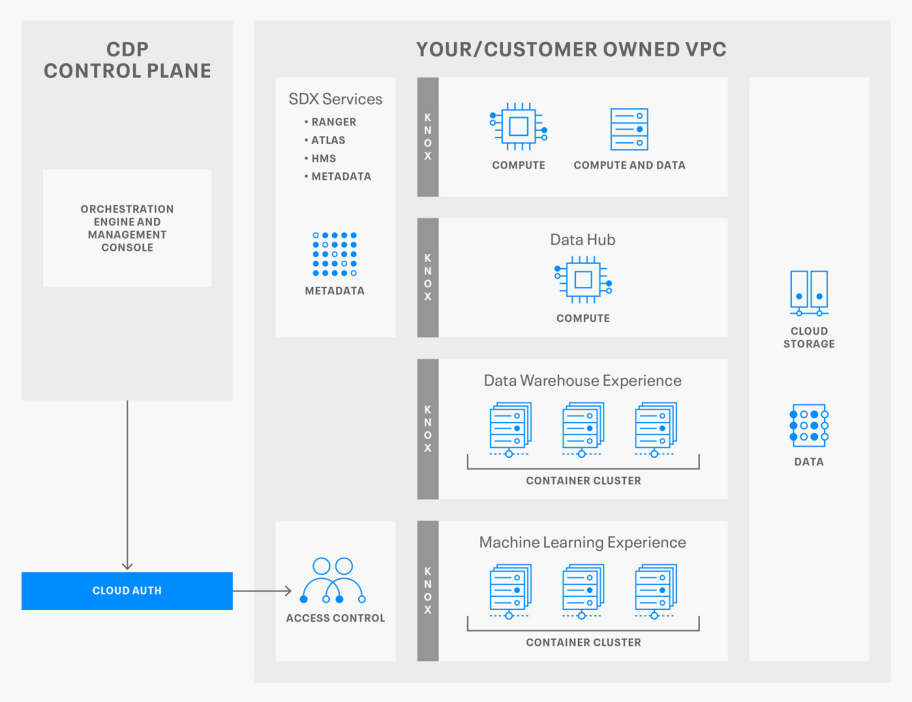



Image 3: CDP Public Cloud Architecture

## Authentication

Within the platform, CDP uses several methods to seamlessly authenticate users across its various components. The Identity and Access Management service (IAM) is responsible for authenticating all calls to the CDP Control Plane and provides authorization support for CDP Control Plane services. IAM does this by federating against your own identity provider via SAML. IAM then uses FreeIPA, Kerberos, and Apache Knox behind the scenes, to ensure that users can access the CDP experiences and clusters without the need for maintaining separate sets of credentials. All communication and authentication between infrastructure within your environment is protected via TLS, or SASL using GSSAPI (Kerberos). If cloud administrators want more granular control over their users' access to data, they can use ID Broker, available in each CDP environment, to map users or groups to public cloud account roles. IDBroker is a feature of Apache Knox, which allows the easy mapping of LDAP users or groups to public cloud account roles, thereby greatly reducing maintenance overheads.

### Encryption

Encryption at rest is enabled on many services by default, such as object storage and cloud native databases (e.g. RDS). Some experiences, such as Cloudera Data Warehouse (CDW) even encrypt attached volumes by default. Services that don't encrypt as standard still have the option available, such as Data Hub, giving you more flexibility to leverage systems and approaches already put in place by public cloud providers.

### Compliance

CDP Public Cloud is SOC 2 Type II compliant. This certification helps ensure that CDP on public cloud is developed, reviewed, tested, and released following the AICPA Trust Services Principles. For InfoSec teams, this means that the CDP Public Cloud service is continuously being developed using audited processes and controls to ensure the highest level of trust and security. To achieve our SOC 2 Type II certification, we demonstrated to our auditors that CDP Public Cloud has suitable policies and controls in place, including:

- A secure software development lifecycle
- Access control that follows "least privilege" best practices
- Detailed logging, monitoring, and alerting
- Encryption controls that meet or exceed best practices
- Completion of internal and external penetration testing
- Active monitoring for intrusion events and security incident handling
- Data backup and disaster recovery

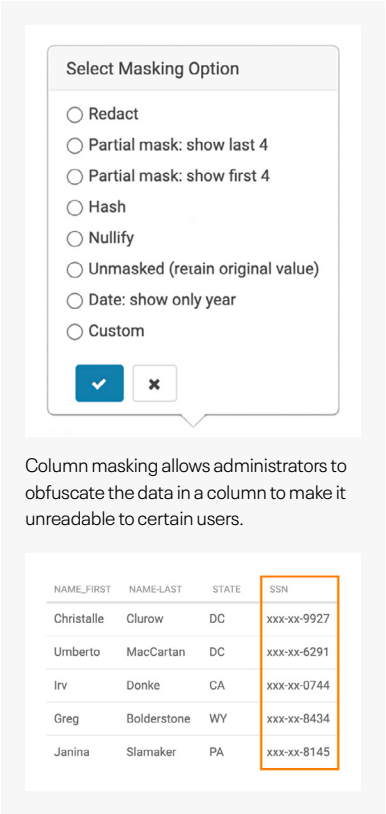
In addition to SOC 2 Type II, Cloudera is working aggressively on further compliance achievements, including expanding Cloudera's ISO27001 certification to include CDP Public Cloud, FedRAMP, and more.

## Cloudera Shared Data Experience (SDX)

As hybrid and multi-cloud landscapes have become the norm for many organizations, so have the challenges of keeping data security and governance policies consistent between different deployments. Implementing and synchronizing policies between public clouds and data center can result in a tremendous amount of operational effort. These efforts delay delivering access to data and analytics for end users and create potential security risks while trying to establish and maintain compliance.

Cloudera's Shared Data Experience (SDX) addresses just those challenges for all CDP deployments. Provided as a seamless integrated data context layer, SDX delivers transparent data security and governance policy management as well as enforcement. Administrators set policies once and have them consistently applied everywhere without additional effort, enabling safe, secure and compliant end user access to data and analytics.

Of the many powerful features in SDX, let's start with the simple column masking example. In many instances administrators need to limit the visibility of a column to protect sensitive data. Normally, a view is created that omits the sensitive data and a user or group is granted access to that view. However, rather than generate separate views based on different combinations of column visibility, it is more efficient to hide or obfuscate a column for a specific user or group. Column masking allows administrators to modify the data in a column to make it unreadable to certain users. Security administrators can choose from a range of pre-built or custom masking options to obfuscate data dynamically. Furthermore, because of SDX, this only has to be set once in Apache Ranger, which is running in your public cloud account, for the policies to then be applied across the entire platform.




Column masking allows administrators to obfuscate the data in a column to make it unreadable to certain users.

Another feature of SDX that is traditionally resolved with the use of views, is the ability to protect ranges of data using a filter statement (similar to a *SQL WHERE* clause). For example, an administrator may want to regionally limit data so that call center employees can only see information from customers in their assigned country or state. Once again with SDX, a security administrator can assign a filter to a group containing users from a geographical area, to limit the accessible data to their region.

Next, we will take some time to talk about another powerful SDX feature, Attribute Based Access Control (ABAC). Many administrators and data stewards may be familiar with using traditional role based access control to grant or deny access to resources. In contrast, attribute based access control allows administrators to grant or deny access to resources through data asset tags or classifications. To help explain this concept some more, let us use the example of a table containing employee data that is used by both HR and a data science team. To hide sensitive data from the data scientists, a traditional role based access control scheme would use two views. One view would hide the sensitive columns, such as salary, while the other utilized by HR would include all of the columns. Creating views to hide different combinations of columns as we add more users or groups quickly leads to the exponential growth of the number of views (a.k.a. view proliferation). Administrators may end up managing dozens of views on a single table, creating more opportunity for costly mistakes. With ABAC, administrators and data stewards simply tag data assets (e.g. Table, Column) with the appropriate classification in Apache Atlas (also part of SDX) and create a tag based data access rule in Ranger. Based on that rule, the tagged asset will be nullified, hashed, or whichever pattern is desired. Furthermore, classifications propagate along data lineage.

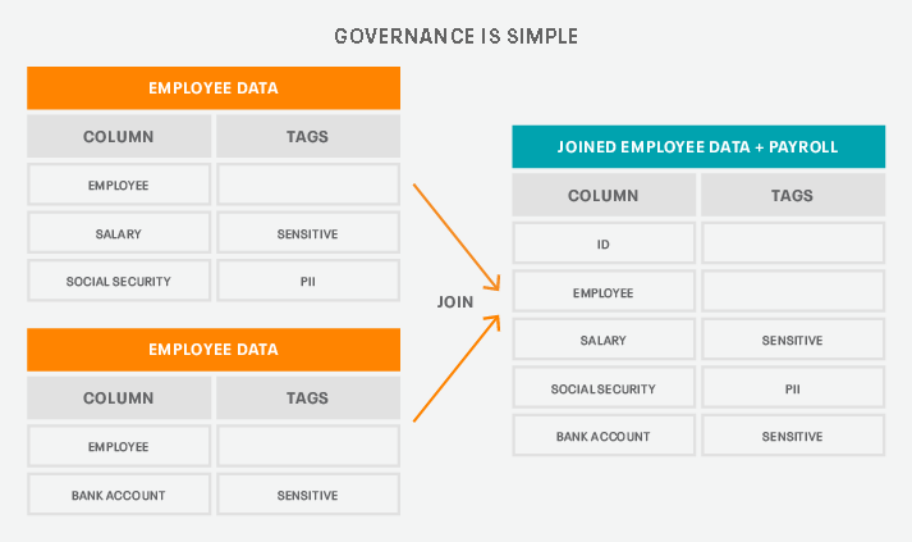



Image 4: Attribute Based Access Control (ABAC) in SDX

When the table containing the tagged data is used to generate another table, the tags will follow. Pulling from our example earlier, if the employee table is joined with another table, the new combined table will inherit the tags from the parent tables; including the tag on the salary column.

With SDX the process of providing temporary access to data has been streamlined by allowing admins to add a validity period to all policies. When creating a data access policy (i.e. basic role based, ABAC, column masking, row filtering, etc.), a validity period can be attached.

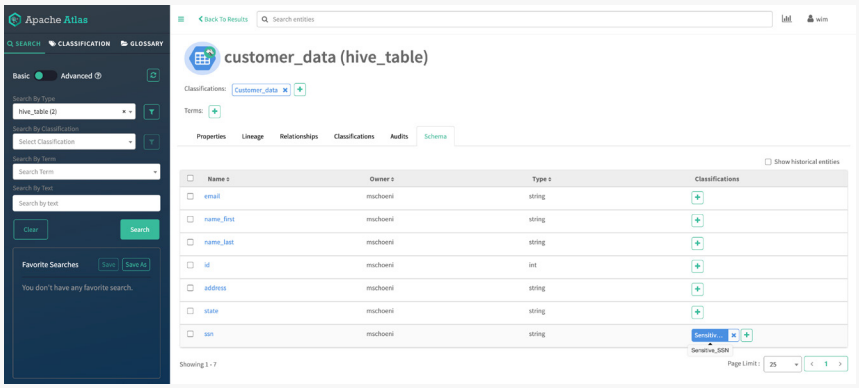



Image 5: Classifying columns based on their content

A further useful feature of SDX is the ability to provide temporary access to data. This task can be a tedious process that must balance security with an administrator's valuable time: time needed to allow and retract access. With SDX, the process has been streamlined by allowing administrators to add a validity period to policies. When creating a data access rule (i.e. basic role based, ABAC, column masking, row filtering, etc.), a validity period can be associated. In fact, more than one validity period can be added to a policy. Recurring access is not supported; a 'set and forget' approach may actually lead to a missed retraction of access and introduce security risk and compliance issues. Now when an administrator creates a policy with the intention of only providing temporary access to a user or group, they can set the start and end times in advance, rather than having to wake up in the middle of the night to grant or revoke access.

We could write an entire white paper on all of the features that SDX has out of the box. Yet one aspect we need to highlight, and where the complete capabilities come together, is the Data Catalog. CDP provides this as a single pane of glass for data stewards and users alike to explore all of your organization's data assets spanning across the platform and its various experiences. The Data Catalog is effectively a front end, with federation capabilities, for the Atlas service running within each of your CDP deployments. Just like Ranger, the Atlas service runs in your public cloud account. Therefore, users can directly access the Atlas user interface for each environment. The Data Catalog is a straightforward mechanism to access the separate instances of Atlas on all of your deployments in one place. This allows you to search across the data assets in all of your connected public cloud accounts. Data Catalog captures the metadata only, not the actual data itself; we ensure that stays securely within your public cloud account. Performing data steward tasks for your organization from the Data Catalog's single pane of glass saves a tremendous amount of time and is significantly more straightforward through the Data Catalog service rather than accessing each instance of Atlas running on each connected public cloud account individually.


Image 6: Data Lineage

**About Cloudera**

At Cloudera, we believe that data can make what is impossible today, possible tomorrow. We empower people to transform complex data into clear and actionable insights. Cloudera delivers an enterprise data cloud for any data, anywhere, from the Edge to AI. Powered by the relentless innovation of the open source community, Cloudera advances digital transformation for the world's largest enterprises.

Learn more at cloudera.com

## Conclusion

CDP's public cloud architecture, together with SDX, ensures deployments are secure by design. CDP's Platform-as-a-Service (PaaS) architecture implements industry best practices for comprehensively secured deployments each and every time, while the SDX data context layer ensures consistent data security and governance as data is used. Together, they give you the flexibility needed to make data available to your end users without compromise, whilst taking full advantage of the agility and elasticity of the public cloud.

Security is a key part of all initiatives to drive insight from data with each organization having unique requirements. To discuss your specific needs in detail, please get in touch with your Cloudera account team.

CDP public cloud is constantly evolving in order to keep the platform secure while providing more important functionality. For the latest information on security and compliance for CDP Public Cloud please visit CDP Trust Center.



Cloudera, Inc. 5470 Great America Pkwy, Santa Clara, CA 95054 USA cloudera.com