## Executive Summary

The purpose of this document is to consolidate information about the Enterprise Data Architecture of Chewy’s Human Resources (HR) organization, identify the current state of data and define a future state and roadmap for HR.

Chewy’s HR organization has invested in and identified data as a key enabler to the growth of Chewy. The Data Enablement tower within Enterprise People Analytics is tasked with making our data easier to use and decreasing the time to value and insights. To achieve this, we need to create an Enterprise Data Architecture that is inclusive of all organizations, scalable and set up for future growth, secure, governed and has strong data quality so that we can trust the data.

*To support Corp, CS, FC in Attrition, DEI, Compensation, L&D, Employee Relations, Compliance, etc. we have a lot of work to do, and the team the roadmap would be quite long.*

*To this end, we are requesting the following additional investment:*

* *Contingent Labor: ~ $300K (Data Architect/Modeler + 2 Data Engineer/Analysts)*

## Background and Assumptions

We define Data Architecture as “…the models, policies, rules, and standards that govern which data is collected and how it is stored, arranged, integrated, and put to use in data systems and in organizations.”[[1]](https://en.wikipedia.org/wiki/Data_architecture)[[2]](https://www.bmc.com/blogs/data-architecture/#:~:text=Data%20architecture%20is%20a%20framework,foundation%20of%20any%20data%20strategy.)This is inclusive of the Data Security and Governance, Data Modeling, Data Quality, Data Definitions, Standards and Metrics in the HR domain.

There has been much work to stand up the initial HR Data Mart which is implemented in Snowflake, and which supports several tableau dashboards. However, the Data Mart in its current form is hard to access and use, and outside the HRIT team there are no other products using this data. The intent behind the data mart is a scalable, one stop shop for all HR data in a consumable, standardized, secure and governed environment. There are many opportunities to scale the Data Mart to serve all of HR.

HR data needs to be properly secured, governed, and handled, and to that point, Chewy HR has made the decision to stand up our own Snowflake instance that is separate from the rest of Chewy data. By creating our own instance, there is additional work that must be done to administer and support the instance apart from the corporate instance.

In terms of the larger Data Architecture, there is an assumption that Workday will be the source for all HR data, but there are other systems (e.g., Greenhouse, Kronos) that also are sources of truth for specific elements of data, and which are supported by different teams. To have a holistic picture of the architecture, we’ll need to account for all the systems and have an enterprise view of the HR Data Ecosystem.

## Current State Assessment

**Source Systems**

Current Workday Integration Architecture is shown in Appendix 2. HRIT owns all the Integrations involving Workday. There is additional documentation directly from Workday in the Resource Center [[3]](https://doc.workday.com/). Additionally There are also integrations between Greenhouse and some vendors., there is some documentation from Greenhouse [[4]](https://developers.greenhouse.io/harvest.html)[[5]](https://github.com/grnhse/greenhouse-api-docs). Greenhouse integrations are owned by the TA Ops team. Kronos is owned by the HR Ops Team

Documentation is scarce on all sides and is not consistently captured. Requests for information also go through each team and causes unnecessary context switching. The security and governance models are defined independently by each team, so while an end user might not have access to a particular piece of data in one system, they can go to another system, and they might be able to get access to that attribute via different access paths.

There is no standard access and governance framework to track what access each person or role has access to. Additionally, while there are data classifications and general data privacy information published by security and compliance, there is no central repository of how the data within each source system is classified, nor is there much guidance on how to properly handle different classifications of data within the HR realm.

**Opportunity**: Set up an enterprise view of the HR Data Ecosystem, inclusive of Workday, Greenhouse and Kronos in standardized documentation and feeding an information catalog.

Create guidance and structure around the classification of data attributes and how to properly handle that data (e.g., what is appropriate to be able to download via Tableau vs what needs to only be allowed in aggregate).

**Reporting**

There are xxx reports that are generated out of Workday, xxx generated out of Kronos, xxx generated out of Greenhouse. Each system has its own team that manages admin and security models for those canned and custom reports. It is unknown how many of these reports are then further enriched, analyzed and then sent on to other people who might receive them, whether in Tableau, Excel or another tool of choice.

In Tableau, there are at 6 different subfolders within the HR parent folder with inconsistent levels of permissions, and another set of reports within the FC parent folder as well. It is not well governed on who has access to each report, and whether those permissions are necessary and appropriate. Details about the known Tableau Reports is attached in Appendix 3.

The way that Tableau is set up across Chewy, the permissions are applied at a folder level, and the assumption is that reports within a folder should have the same security model. However, in the HR space, different teams have created different folders with different strategies, so the permissions are managed currently by department or group, rather than logically by what a role should have access to. This leads to a poor user experience, where a user needs to request access to each department folder individually, and often may get access to data they don’t need nor care about.

A lot of reports in tableau are curated to only display access to aggregated data, but the raw underlying data is at a person level and has lots of detailed information that is not necessary to the report. Because of the permission by folder level and many people being in the Explorer category, those granular level details of data are available via download in the Tableau report, and we lose sight of what a user does with that data downstream.

The only automated Tableau Dashboards are the ones built or inherited by the HRIT team and use the HR Data Mart as a data source (Project Viking, Eagle Eye, HRBP, etc.). Other Tableau Dashboards that are not automated and not using the HR Data Mart are using manual data refreshes (i.e., Flat File Ingestion) are published to the development server (reports.chewy.com).

**Opportunity**: Find and document the “Menu of Reports” and have standard reports for each purpose to ensure standardized reporting capabilities all around. Redesign and govern access to HR Tableau reports in such a way that the Tableau is secure and governed, while having a pleasant user experience. Automate core reports from data out of HR Data Mart and continue to grow the Data Mart to be more inclusive of all the asks.

We also need to design and publish a strategy on when to use aggregate data for sensitive data, and when granular level details are ok. This blend of aggregate and sensitive will be particularly relevant in the HR data domain.

Additionally, we should be able to design the Tableau folder structure for HR in a way that is driven by access and role, rather than by individual folders. We should also develop a comprehensive strategy for when to use aggregate and when to use granular data in Tableau, especially for our sensitive data.

**Snowflake**

**Technical Capabilities**

HR data is inherently sensitive in nature as it deals with PII and employee data. Consequently, HR has decided to create an instance of Snowflake that is separate from the rest of the Corporate Chewy Snowflake instance. The HRIT team administers and develops code on this Snowflake instance apart from the rest of Chewy. Because we have a separate instance, that is solely administered and developed upon by the same team, there are many steps that are not taken.

|  |  |  |
| --- | --- | --- |
| **Capability** | **HR Snowflake** | **Chewy Snowflake** |
| Continuous Integration / Continuous Deployment (CI / CD) |  | X |
| Code Review and Architecture Approval |  | X |
| Dev / QA Environments |  | X |
| Sandbox |  | X |
| Automated Admin through SNOW |  | X |
| Data Catalog Integration (Mr. Peabody) |  | X |

The Chewy Snowflake environment has defined processes for continuous integration and deployment, which requires code review and approval from the central engineering team to ensure good performance and scalability. It also has development, QA and sandbox environments which the HR Snowflake account has not stood up yet. In turn, the development of all code is done in the production environment in the HR Snowflake account and has immediate impact on anyone using that data.

Administering of the cluster is also done manually within the HR Snowflake account, and permissions are run line by line, not recorded anywhere and hard coded when requested. The Chewy Snowflake cluster has had a role design that has been created and permissions are granted through a standard SNOW process[[6]](https://chewyinc.atlassian.net/wiki/spaces/EDW/pages/6834612/AD+to+Snowflake+Role+Mapping).

Finally, the Chewy Snowflake account is tied into the Mr. Peabody Data Catalog, so the data is documented, and usage can be tracked via the tool. HR has deliberately chosen not to integrate with Mr. Peabody as HR metadata itself can be sensitive, but in doing so there is a lack of solid documentation and no standard repository to collect that metadata.

**Opportunities:** There is a lot of opportunity to learn from what the Enterprise Chewy Snowflake team has built and use their expertise on the system architecture, process and strategies. There’s no need to re-invent the wheel, the HR Snowflake Cluster is lacking in scalability.

Specifically, we should enable CI/CD and a Development Process (e.g., Code Repositories, Code Reviews, Automated Deploys, Development and QA Environments) and create an access strategy that is not ad-hoc, and follows the various layers of access and security that have been designed at the source.

There is also an opportunity to either have our own Mr. Peabody instance or create our own centralized Information Catalog.

**Data Model**

The Data Model in Snowflake was created on a project-to-project basis. Views are created with business logic hardcoded in the view, instead of creating generic views that can be re-used across various analytics requests. In addition, the data model is created in a way such that access to data is granted in an all or nothing fashion. Details can be found in Appendix 3.

Views are complex and have a lot of logic embedded in the views, and logic is reused across multiple views. When changes are needed, those updates are needed across all the views to keep them consistent. However, since there is no deployment pipeline, nor is there a dev environment, the changes are done to the production data model. This seemingly happens quite often in particular around mappings and hierarchy information.

This type of reference and master data is an area of opportunity to gain some efficiency. We should standardize reference data across our source systems (e.g., Workday, Greenhouse, Kronos) and our partner source systems as well (e.g. Hyperion).

An example of a proposed Curated and Business Product Layer is shown in Appendix 4.

**Opportunity:** If we focus on designing and building a Curated and a Business Product Layer that is scalable, it will be able to serve all reporting and analytic needs without constant manual intervention.

To do this, we need to better understand ourbusiness processes and capabilities, and design by domains, rather than by project.

**Historical Data**

In reporting and analytics, historical or point in time data is necessary to do trending analysis. In the current HR Data Mart, data is pulled in each day to show what is stored in the system for current state, however, the data of what data looked like in history is overwritten and lost. Not having the history can lead to confusion when comparing metrics across time periods and the data looks like it has changed.

**Opportunity:** Snapshots of key data and metrics would be advisable.

**Architecture**

There are lots of pieces to the Data Architecture of the HR Domain. Each individual piece of integration and team may have their own documentation, but there is not standard documentation nor a consolidated view of the Architecture from a holistic point of view. To obtain a holistic picture of the HR landscape, we need to know what business domains and capabilities we have and what data is created in each of those domains. Not just for the major data sources (e.g., Workday, Greenhouse, Kronos), but also for other data that is typically manually assessed (e.g., Glint, Recruitics, Learning Data, Assessments, OpenBark, etc.).

From the opposite end, we can generate the architecture by profiling the data products that consume the data (e.g., Tableau reports – Project Viking, Eagle Eye, HRBP, etc.) and see what the core data attributes are that are being used. This work has been started. See Appendix 5 for examples of these top down and bottom-up approaches to discovering the current Data Architecture.

**Opportunity:** Create both a business capability model and build documentation of the Tableau reports up to discover what the Data Architecture should be if driven by Data Domains.

To truly understand how data in the HR domain is accessed and governed, we need to know what data products exist. From that list, we can design proper access, governance, standards and definitions. A data product in this context can be

* Source Systems (Workday, Greenhouse, Kronos)
* Operational Reports
* Historical or Analytic Reports
* APIs
* Data Science Model
* Metric or KPI

In terms of security and access governance, as has been stated, there is not a consistent process nor guidance on how to handle the sensitive data. We should focus on creating that process with documentation and training on how to safely handle our data products. In the current state, all products should go through the PII request process to be cleared, but am unsure how often that is happening.

**Opportunity:** If we compile a central repository of Data Products, this will be a useful tool across the entire HR Data Landscape. In compiling this list, we gain documentation and insight into what exists, but also are then able to coherently design access and governance strategies. Additionally, all data products should go through the legal and compliance review process that we define and document. At a minimum it should go through an inform for any product that does not use sensitive data.

Another thing that is missing from our current architecture is a common agreement and alignment on business roles. Once we can come up with standard business roles, we can apply those roles and appropriate access to data at the role level (role-based permissions). This will allow more comprehensive audits and we can find where we need to fill the gaps and create more access controls and governance in the different source systems. As of current state, these business roles do not exist, and permissions are given on an ad-hoc basis and are cloned from person to person (e.g., “Can you grant me the same levels of access that my coworker has”). This can be dangerous as people inherently should move between roles over time, and their access should also change with the role, rather than gaining access to sensitive data and then being grandfathered into that access group when it’s inappropriate to continue to have access to the data. The level of governance now is REACTIVE to changes, and are done only on periodic audits, rather than being alerted of a change in job status.

**Opportunity:** We should identify the various personas in and across Chewy and map them to appropriate access to HR data attributes and products. From that map, we can develop a governance and alerting framework for when people move between roles and no longer need access, and grant access based on roles, instead of manually assigning in an ad-hoc fashion.

One way to ensure appropriate access to the data, we need to handle data appropriately as we develop for both our source systems and also our reporting systems. In our current state, all development is done against real employee data, and tableau reports are all built off row level granular data. To ensure proper access to data, and in the spirit of only accessing the data necessary to do one’s job, we should have test data profiles set up across the various systems, and also a way to anonymize data in our test environments to do true testing while not opening ourselves to risk of real production and person level data.[[7]](https://www.adessagroup.com/hr-gdpr-cloning-scrambling-and-anonymization-of-employee-data-in-sap-hr/)

**Opportunity:** Using either a data anonymization tool, scrambling the data, or simply to create test subjects across our source systems will allow us to do much more targeted testing while not risking production, employee level PII. This test-driven development would allow us to more effectively and efficiently configure both source systems, and our downstream HR Data Mart and analytics.

Lastly, business rules are not defined in a standard way across the HR ecosystem, and are defined on an ad-hoc basis. Of the business rules that we do have and have defined, there is very little Data Quality checking that is done outside of manual processes and human intervention. This is a cause for context switching and is error prone, leading to inefficiency in the communication cycle. By discovering existing business rules, defining new ones and capturing all business rules in a central place, we can put a Data Quality framework in place that will help Create Trust in the HR data in the key dimensions of Data Quality[[8]](https://www.collibra.com/us/en/blog/the-6-dimensions-of-data-quality)

* Completeness – Is all the data being correctly captured, and stored in our Source Systems and in the Data Mart?
* Accuracy – Is the data accurately representing the HR processes that we have put in place?
* Consistency – Is the data consistent across Workday, Greenhouse, Kronos, Snowflake, Finance, etc.?
* Validity – Is the data valid for the different use cases (e.g., picklists, dropdowns, departments, etc.)?
* Uniqueness – Are there duplicates in the data either in each system and across systems?
* Integrity – How does the HR data move across different systems and can it be traced?
* Timeliness – Is data getting where it needs to go in a timely manner which meets SLAs?

**Opportunity:** Create a Data Quality framework that tracks and remediates all data quality issues across the HR data ecosystem.

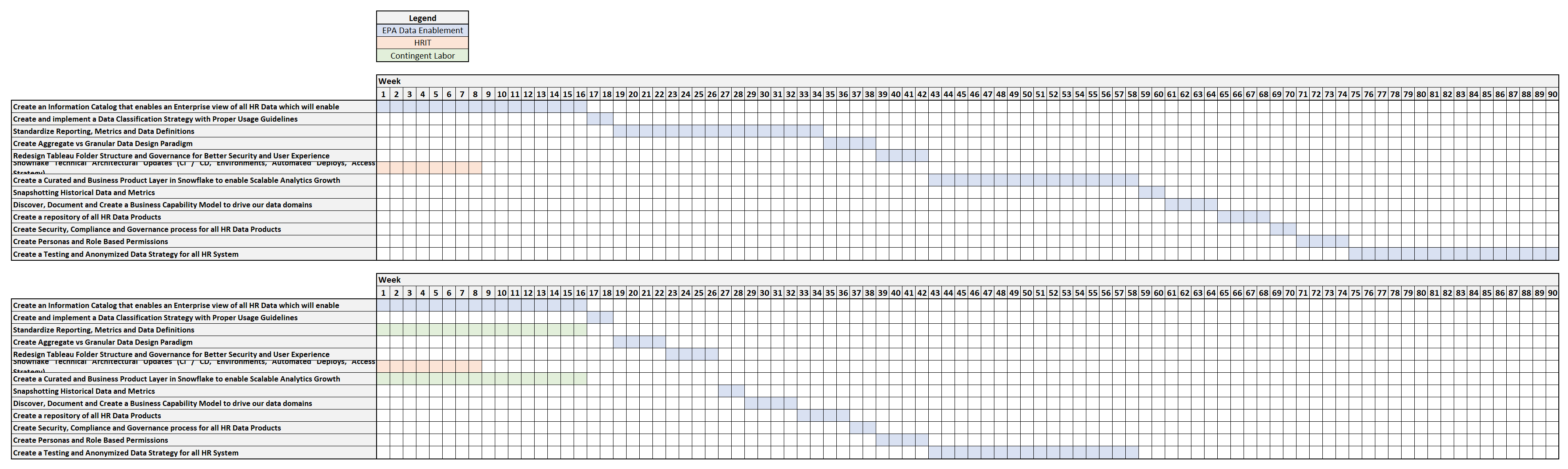
## Future State and Opportunities for Growth

We have identified the various opportunities as we have stated in the Current State Assessment above, but in summary, here are the various Opportunities that we have for HR Data Architecture. By implementing these opportunities, we would enable data across HR and help HR continue the desired path to being a World Class HR organization.

* Create an Information Catalog that enables an Enterprise view of all HR Data which will enable
  + Downstream Impact and Upstream Source Traceability
  + Better Access and Governance of Data
  + Central Metadata Repository of all HR Knowledge and Information
  + One-stop shop for HR Information
* Create and implement a Data Classification Strategy with Proper Usage Guidelines
* Standardize Reporting, Metrics and Data Definitions
* Create Aggregate vs Granular Data Design Paradigm
* Redesign Tableau Folder Structure and Governance for Better Security and User Experience
* Snowflake Technical Architectural Updates (CI / CD, Environments, Automated Deploys, Access Strategy)
* Create a Curated and Business Product Layer in Snowflake to enable Scalable Analytics Growth
* Snapshotting Historical Data and Metrics
* Discover, Document and Create a Business Capability Model to drive our data domains
* Create a repository of all HR Data Products
* Create Security, Compliance and Governance process for all HR Data Products
* Create Personas and Role Based Permissions
* Create a Testing and Anonymized Data Strategy for all HR System

## Roadmap

Details of how this roadmap is calculated can be found in Appendix 6



Without any contingent labor, the Enterprise People Analytics Data Enablement team has the roadmap marked out to take 90 weeks.

With contingent labor added, the roadmap shrinks to 58 weeks, and delivers the Curated and Business Product Layer for the Data Mart as well as standard reporting, metrics and definitions in the first 16 weeks.

**Note:** With contingent labor, Enterprise People Analytics Data Enablement would still drive the strategy and architecture, and the contingent labor would work directly with the EPA team. Work products will be created without the long term need for support.

## Next Steps

Next step is to decided if we want to invest in the contingent labor option of $300K.

If contingent labor is not approved, the Enterprise People Analytics Data Enablement team will prioritize, and start to work on the 90 week.

If contingent labor is approved, we will begin sourcing immediately and begin work on both the contingent labor ask, as well as prioritizing the remaining work.

NOTE: These timelines are based on HR, IT, and Product alignment for the HR product approach outlined in the document and resource availability (to be confirmed), therefore these timelines will be finalized once alignment and resources are confirmed.

### APPENDICES

### Appendix 1

**Reference Links**

[1] <https://en.wikipedia.org/wiki/Data_architecture>

[2] <https://www.bmc.com/blogs/data-architecture/#:~:text=Data%20architecture%20is%20a%20framework,foundation%20of%20any%20data%20strategy>.

[3] <https://doc.workday.com/>

[4] <https://developers.greenhouse.io/harvest.html>

[5] <https://github.com/grnhse/greenhouse-api-docs>

[6] <https://chewyinc.atlassian.net/wiki/spaces/EDW/pages/6834612/AD+to+Snowflake+Role+Mapping>

[7] <https://www.adessagroup.com/hr-gdpr-cloning-scrambling-and-anonymization-of-employee-data-in-sap-hr/>

[8] <https://www.collibra.com/us/en/blog/the-6-dimensions-of-data-quality>

### Appendix 2

**Current HR Integration Architecture Diagram** *(from Workday perspective)*

Diagram

Description automatically generated

**Source:** <https://chewyinc.atlassian.net/wiki/spaces/HRIT/pages/1437336102/Current+HR+Systems+Architecture+-+January+2022>

### Appendix 3

**HR Data Mart Details**

Appendix 4

**Proposed Curated and Business Product Layers in HR Data Mart**

### Appendix 5

**HR Data Architecture Discovery**

### Table Description automatically generated with medium confidence

A picture containing table

Description automatically generated

### Appendix 6

**HR Data Architecture Roadmap Details**

Small = < 2 Weeks

Medium = 2-4 Weeks

Large = 4-8 Weeks

X-Large = 8-16 Weeks

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Opportunity** | **Category** | **T-Shirt Size** |
| 1 | Create an Information Catalog that enables an Enterprise view of all HR Data which will enable | Foundational | X-Large |
| 2 | Create and implement a Data Classification Strategy with Proper Usage Guidelines | Governance | Small |
| 3 | Standardize Reporting, Metrics and Data Definitions | Tableau | X-Large |
| 4 | Create Aggregate vs Granular Data Design Paradigm | Tableau | Medium |
| 5 | Redesign Tableau Folder Structure and Governance for Better Security and User Experience | Governance Tableau | Medium |
| 6 | Snowflake Technical Architectural Updates (CI / CD, Environments, Automated Deploys, Access Strategy) | Data Mart | Large |
| 7 | Create a Curated and Business Product Layer in Snowflake to enable Scalable Analytics Growth | Data Mart | X-Large |
| 8 | Snapshotting Historical Data and Metrics | Data Mart | Small |
| 9 | Discover, Document and Create a Business Capability Model to drive our data domains | Foundational | Medium |
| 10 | Create a repository of all HR Data Products | Foundational | Medium |
| 11 | Create Security, Compliance and Governance process for all HR Data Products | Governance | Small |
| 12 | Create Personas and Role Based Permissions | Governance | Medium |
| 13 | Create a Testing and Anonymized Data Strategy for all HR System | Foundational | X-Large |

A picture containing calendar

Description automatically generated

**Details**

EPA Data Enablement Only – 90 weeks

EPA Data Enablement + Contingent Labor – 58 Weeks

**Note:** Technical Architecture Updates for Data Mart are solely at the discretion of HRIT, but that work has not been prioritized nor resourced for.