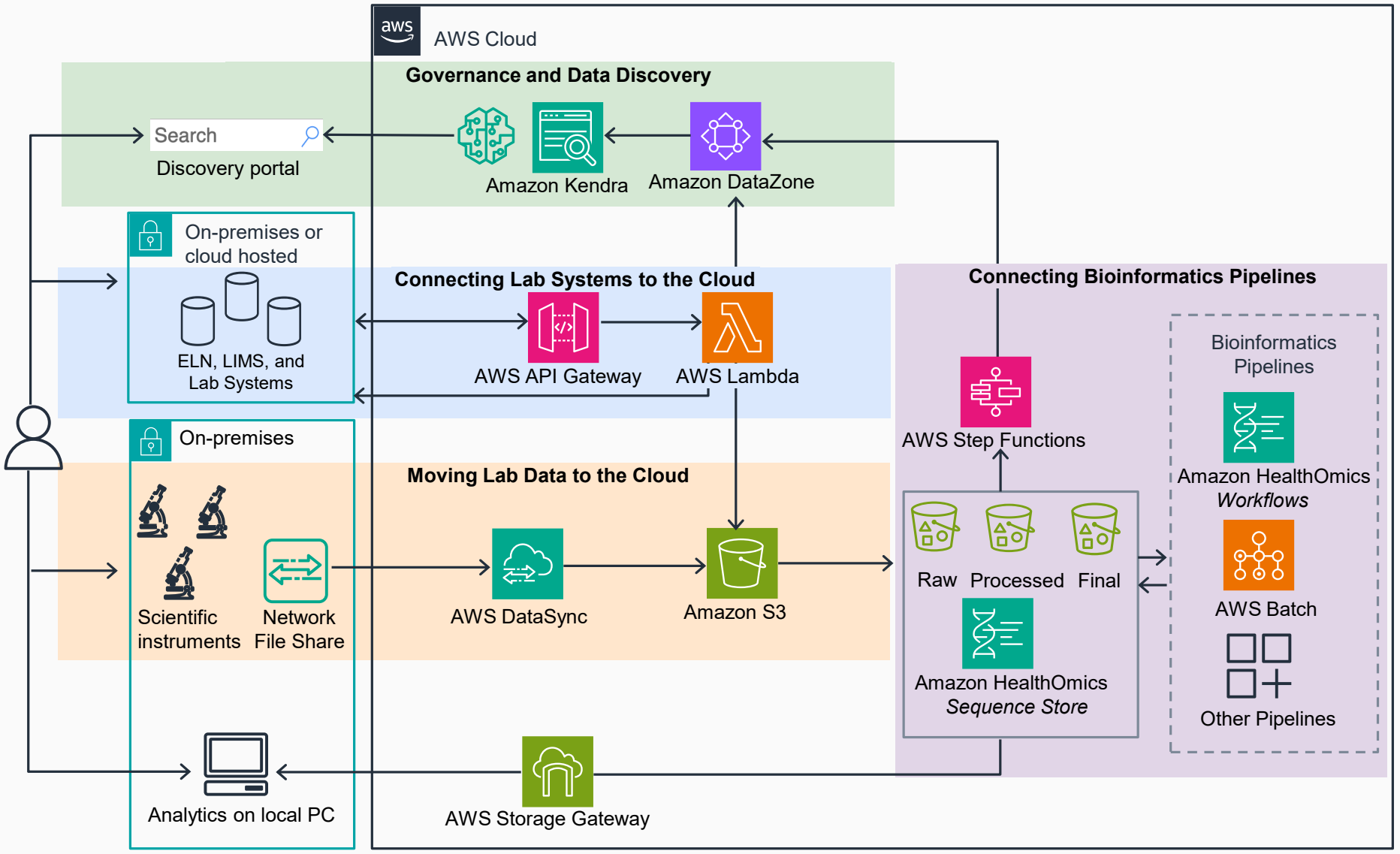
Guidance for a Laboratory Data Mesh on AWS

This architecture diagram shows an overview about how you can accelerate the launch of a scientific data management system that integrates both your laboratory instruments and software with cloud data governance, data discovery, and bioinformatics pipelines, capturing key metadata events along the way. For more details on each component, follow the next slides.



Goal:
This Guidance provides details on the implementation best practices for metadata enrichment, search, and bioinformatics processing as part of Building Digitally Connected Labs with AWS.

Conceptual overview:
When a scientist or technician sets up an experiment in the electronic lab notebooks (ELN) or lab information management systems (LIMS), this notifies the metadata catalog of that new experiment and configures the data store to receive that instrument data.

Later, when lab instrument data is collected, the data moves to the data store that was pre-associated with the metadata store.

Bioinformatics processing steps and output files are captured within the data store, and all new files are linked to the ELN or LIMS through the metadata Store.

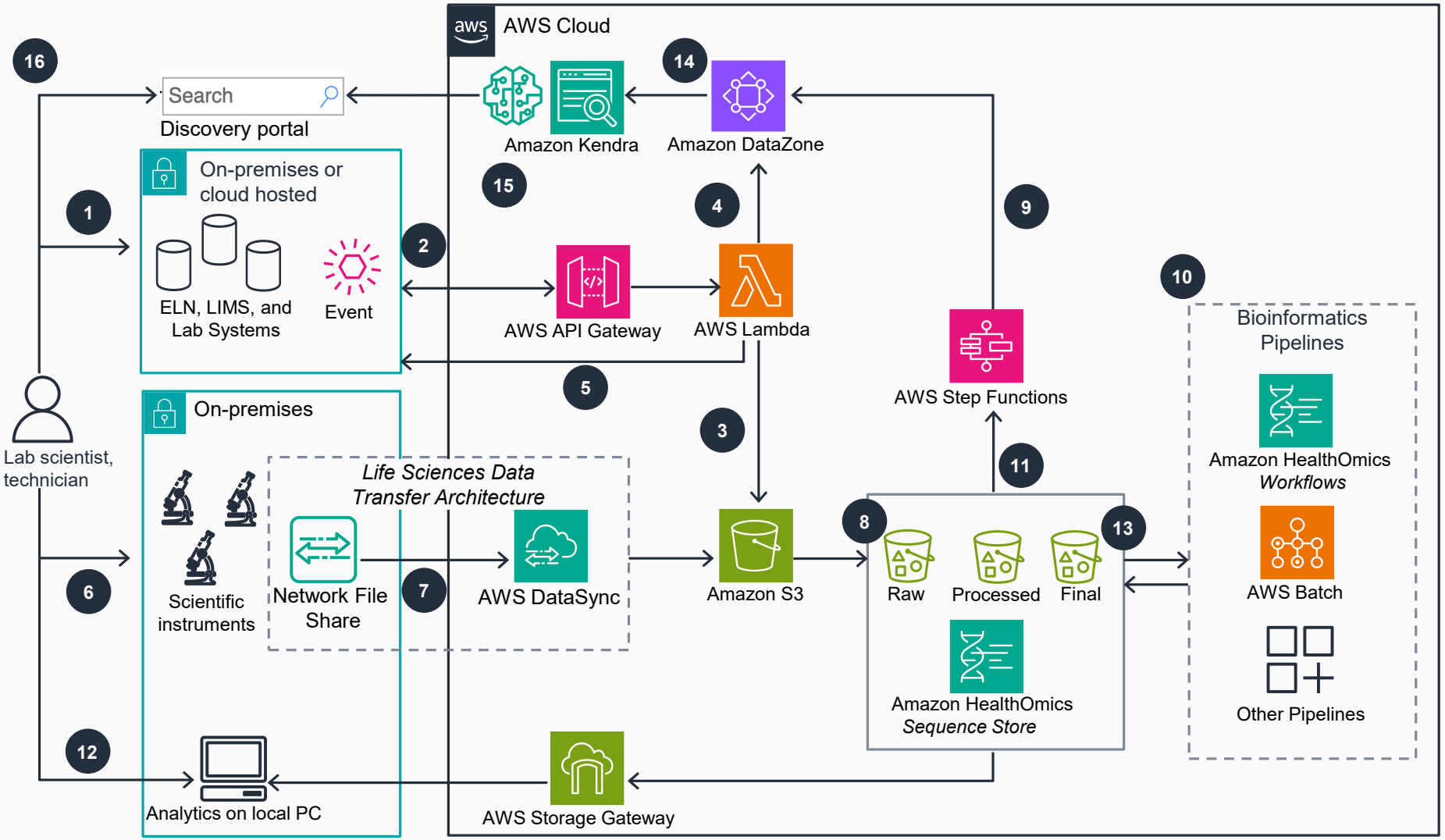
Data is governed and discoverable by metadata search, or by natural language search through a chat interface.

Reviewed for technical accuracy November 13, 2023

AWS Reference Architecture

Guidance for a Laboratory Data Mesh on AWS

Connecting Lab Software to the Cloud. Steps 1-5



Overview: Connecting Lab Systems to the Cloud

Lab software systems like electronic lab notebooks (ELN) and lab information management systems (LIMS) need to have two-way communication with the cloud. This helps with the following:

- To take advantage of scalable storage as an extension of the ELNs.
- For the software to maintain the latest information about new files that are moved to or are processed on the cloud, which may reside outside the ELN workflows.
- For cloud services to have contextual information about datasets for improved data governance and enterprise search.

1 A scientist or technician sets up a project or experiment metadata within the ELN, LIMS, or other experiment or testing database.

2 Upon the creation of an experiment in the ELN, the ELN creates an event that calls an AWS API to send an Experiment ID. With the Experiment ID received, an **AWS Lambda** function calls an ELN's API to retrieve all of the experiment metadata that will be relevant to contextualize the data for discovery.

3 The **Lambda** function sets up **Amazon Simple Storage Service** (Amazon S3) buckets as a scientific Data Store. The setup includes the naming of related folders, based on the unique Experiment ID. At this step, the Data Stores are empty. In addition to this, next-generation sequencing (NGS) data can be stored in **AWS HealthOmics**.

4 The **Lambda** function writes the metadata that it has collected into an **Amazon DataZone** Metadata Catalog. This is done by creating **Amazon DataZone** data assets, adding metadata forms to those assets, and assigning the metadata to those fields.

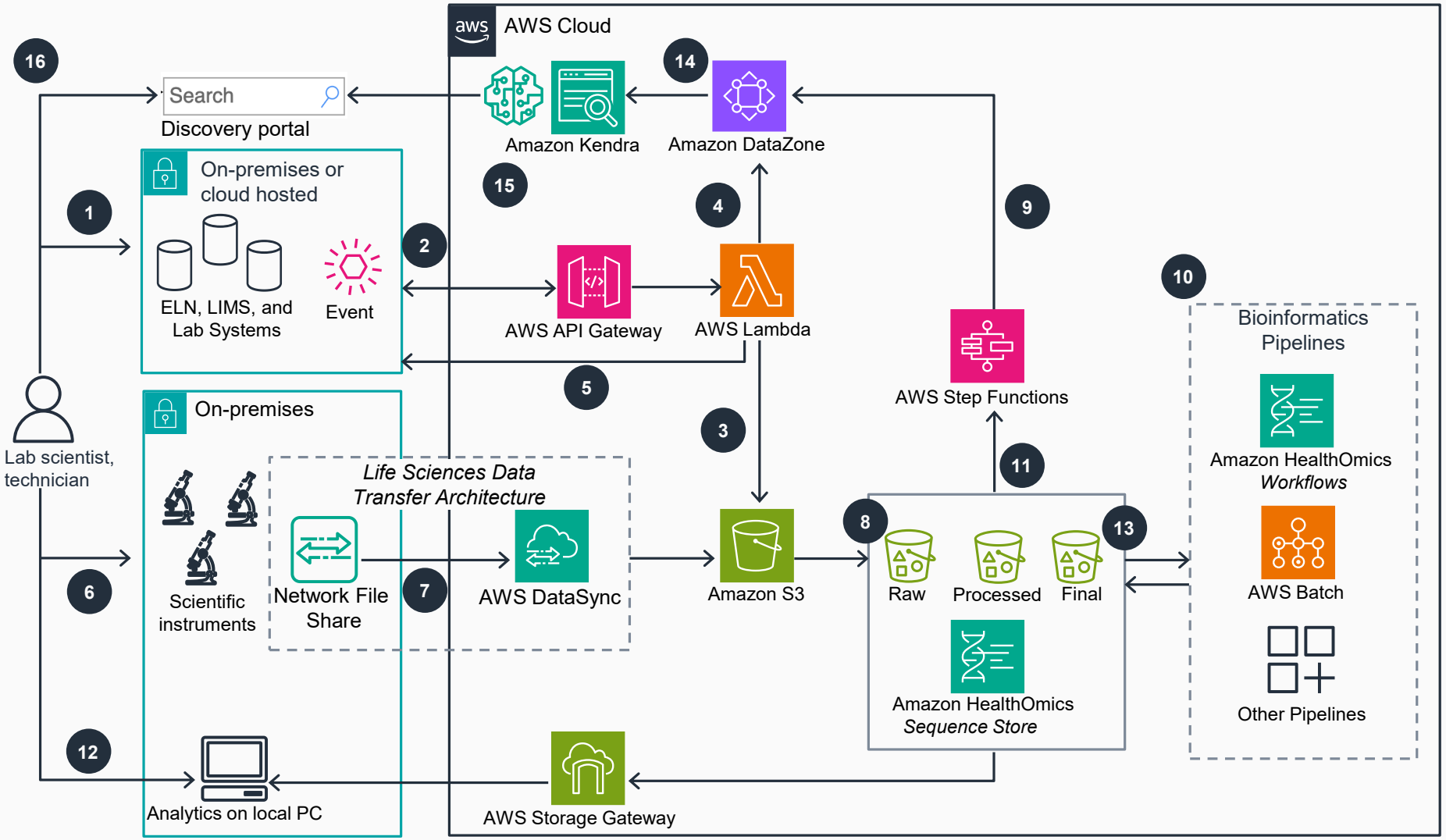
5 **Lambda** calls the ELN's API to add the location of the new data asset to the experiment entity within the ELN.

Reviewed for technical accuracy November 13, 2023

AWS Reference Architecture

Guidance for a Laboratory Data Mesh on AWS

Connecting Lab Instrument Data to the Cloud. Steps 6-9



Overview: Moving Lab Data to the Cloud
Lab instrument data should be automatically moved to the cloud to make it accessible to elastic cloud computing, for data discoverability, and for cost savings. A component of this is outlined in How to move and store your genomics sequencing data with AWS DataSync.

6. Scientists run instruments to collect data and save data to a network-accessible folder being monitored by **AWS DataSync**.

7. **DataSync** ingests the data into the landing zone bucket within the data store.

8. The writing of the file to the data store invokes an event, initiating **AWS Step Functions**. **Step Functions** will import the instrument data from the landing zone into the relevant raw bucket within **Amazon S3**, and optionally into the relevant read set within the **HealthOmics** sequence store.

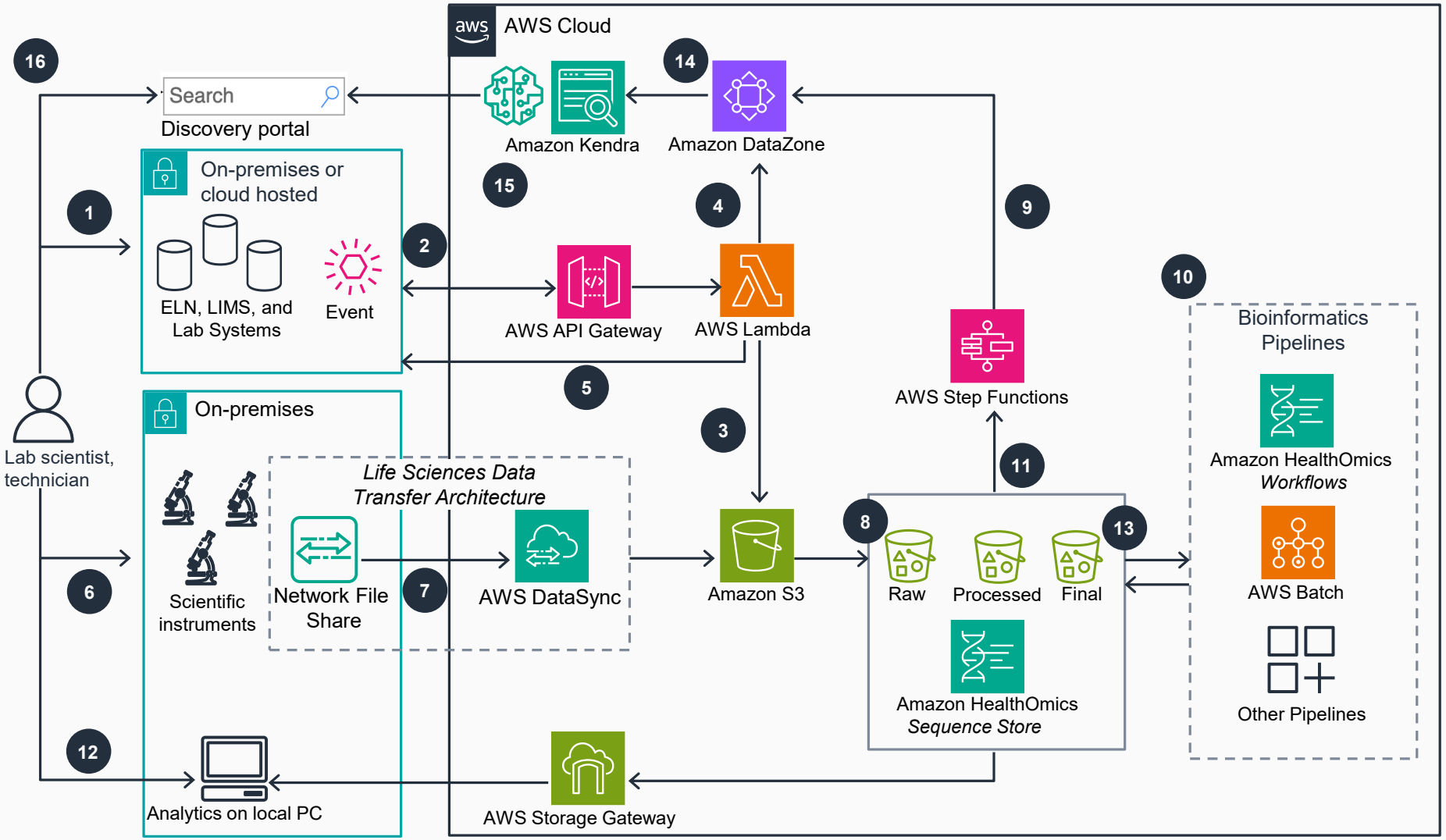
9. **Step Functions** adds the file names, creation date, and other metadata that is extracted from the files to the metadata store for the data assets of the relevant experiment.

Reviewed for technical accuracy November 13, 2023

AWS Reference Architecture

Guidance for a Laboratory Data Mesh on AWS

Connecting Bioinformatics Pipelines. Steps 10-13



Overview: Connecting Bioinformatics Pipelines

At this point, instrument data will be located in the raw bucket, and the Metadata store will have contextual information about that data. Here we discuss how to submit that data to bioinformatics or high performance compute (HPC) processing in a way that maintains association with the metadata store.

10 An **Amazon S3** bucket event initiates a bioinformatics pipeline run using the raw bucket as the data source. Bioinformatics output files are written to processed bucket.

11 **Step Functions** adds the file names, creation date, and other metadata that is extracted from the files to the metadata store for the data assets of the relevant experiment.

12 Optionally, **AWS Storage Gateway** mounts onto the local network for users to access the processed bucket for local analysis or report generation.

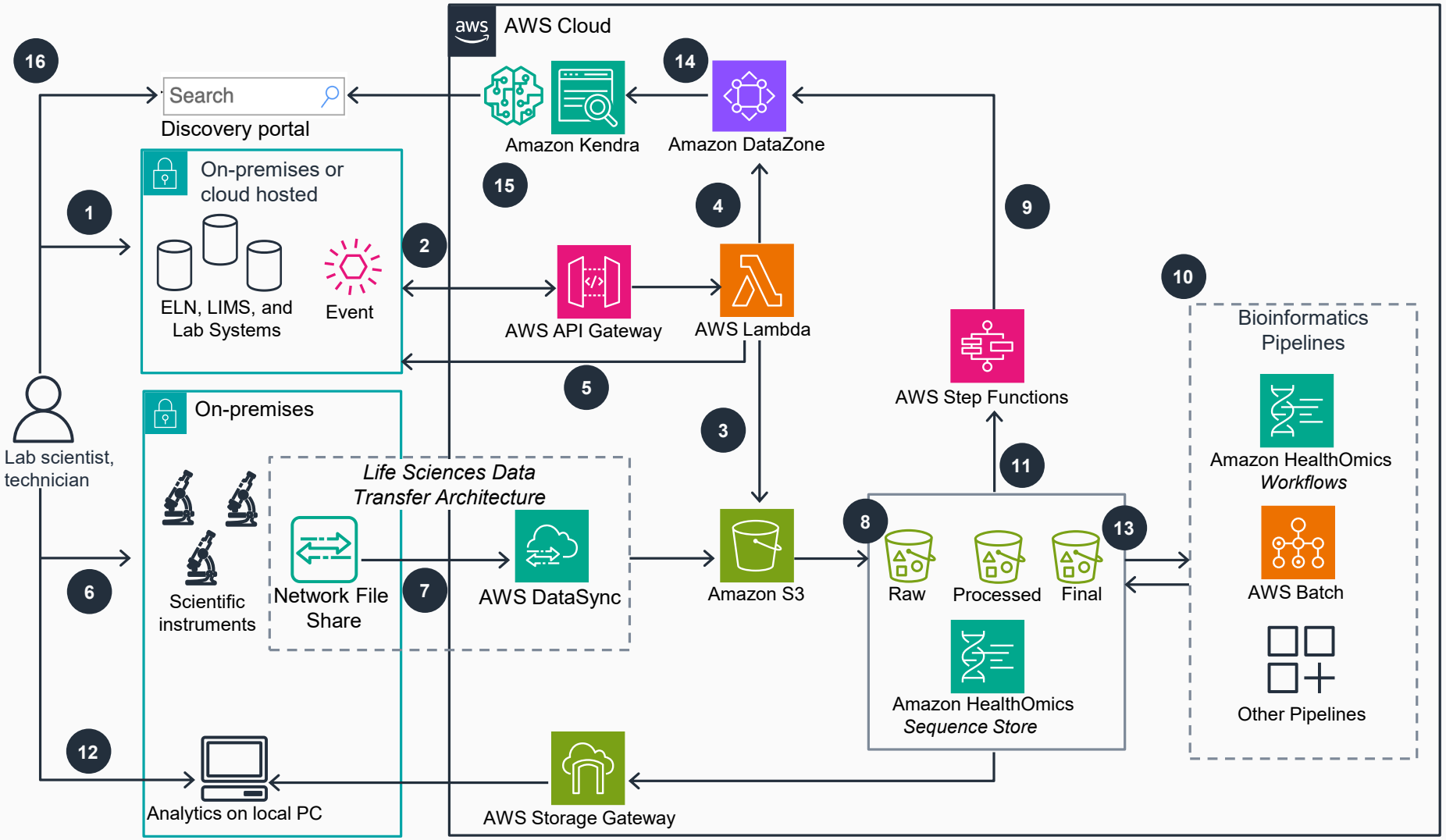
13 Locally-generated files are written to the final bucket. **Step Functions** posts the name and date to the processed data asset in the metadata store.

Reviewed for technical accuracy November 13, 2023

AWS Reference Architecture

Guidance for a Laboratory Data Mesh on AWS

Governance and Data Discovery. Steps 14-16



Overview: Governance and Data Discovery
By this step, instrument data has been collected and associated with laboratory metadata. The lab metadata now provides rich context for the associated data sets and can be used to power a search tool, for lab users to discover, access and use the data. Large language models (LLMs), a class of foundation models (FMs), can also be used in combination with semantic search to create conversational data exploration tools (chatbots).

14 **Amazon DataZone** is a data management portal to discover, analyze, and report on data. In **Amazon DataZone**, research scientists and business users can search for lab datasets using keywords that are found in the metadata store, which originated from the ELN. For example, these may be searches for sample id, experiment id, group, platform, file names, dates, or keywords within the experimental description. These searches will return a list of data assets that have an association with those keywords, which are collections of **Amazon S3** objects.

15 To index the contents of the data files, **Amazon Kendra** is a managed service for indexing data and conducting semantic search. **Amazon Kendra** APIs may be used within a custom discovery portal application that your organization creates, to enable this search. In combination with **Amazon Kendra**, large language models (LLMs), a subset of foundation models (FMs), can be used to generate summaries of search results and create conversational experiences.

16 Data and metadata can be semantically searched by a user to discover datasets, gain access to, and analyze datasets.

Reviewed for technical accuracy November 13, 2023

AWS Reference Architecture