6.

|  |  |
| --- | --- |
| **Dataflow** | This service allows you to manage stream and batch data processing through the running of Apache Beam jobs. It provides a layer of abstraction between the processing logic and the underlying Compute infrastructure to the point where it isn’t necessary to provision infrastructure ahead of time in the same manner that you would on other platforms. In this way, Dataflow fundamentally works on a serverless paradigm and fully takes care of all the management when it comes to scaling your clusters and distributing tasks. |
| **Dataproc** | This is Google’s version of a fully managed service used for Big Data processing applications. While it is commonly used to implement Apache Spark/Hadoop (vs Apache Beam), it is compatible with other similar platforms as well, given that under the hood it relies on Compute Engine instances to run incoming jobs. In addition to this added layer of management abstraction, one of the primary advantage of Dataproc is the fact that it separates data from the processing instance, which is more efficient from a cost perspective as the equivalent server clusters need not be running all of the time. That all being said, it runs “closer to the metal” than Dataflow does, despite being suitable for similar use-cases such as ETL, allowing a more hands-on (server-based) approach to managing the underlying infrastructure. |
| **Dataprep** | The most dissimilar of the three, Dataprep’s main utility lies in its ability to provide visualizations on Big Datasets and perform cleaning/prep operations upon them. It is largely UI-driven, as it is meant to be used in conjunction with some other data source in most cases. As such its ability to comprehensively ingest streams of data is very limited, however it is much easier to mange than the other two options if this is not a requirement. |

7.

|  |  |
| --- | --- |
| The Application | A cloud service used to monitor urban wildlife populations and provide public-sector researchers/organizational users with insights into their intra-urban behavior and population dynamics. The application would consist of a series of IoT Edge clients which push records to a cloud backend when they detect a particular species. The collection of these sightings would then be available for batch processing operations, such as geographical clustering, to extract further insight out of them. |
| Its Impact | Currently in this domain there is a lack of body of knowledge on urban wildlife population dynamics, as the domain is relatively recent. As such, there are not a lot of “tools of the trade” established for practitioners to use at this time. An application such as this would allow its users to accurately assess the likelihood of public activities infringing upon active wildlife territories (and vice-versa) based on geographical location and help implement locally tailored measures that safeguard the public/wildlife populations from unintended interactions with one another. As an added benefit, it would allow project managers/leaders to quickly evaluate project impacts on surrounding ecology (without having to invest in costly comprehensive environmental studies right up front) |
| The Dataset(s) | A good example dataset would be the North American Camera Traps Dataset (<https://lila.science/datasets/nacti>) which contains approximately 3.7 million image records spanning 28 categories/classifications (including “empty”) |
| Pipeline | C:\Users\100613386\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\AB89EB8.tmp |