Intergalactic Cookie Oven Pipeline

Artifacts

Each stage produces artifacts.

Assumptions:

Output of a pipeline e.g. dough created from mixing the ingredients cannot be reversed to get the raw material back.

Dough created by the **mixing** stage is perishable and needs to be consumed by the next stage ASAP.

Storing the artifacts

Based on the above assumption I would store the artifacts of each stage in event hubs.

Raw Material

Some raw materials will be perishable where others won't. The perishable ones will be stored on event hubs and the others in delta lake.

Pipeline Stage Option 1 Databricks

Each pipeline stage will be setup as a job in databricks. This will be stream raw materials / artifacts from previous stages as soon as they become available and pass them on the next event hub.

With databricks we get the advantage of fine tuning the hardware requirement of each stage. If let's say takes more time, we can give it more power by adding additional workers, and bringing down production time.

Each databricks job is agnostic and unaware of the subsequent or previous jobs.

What if databricks cannot handle a stage.

As communication is carried out purely by event hubs we can easily slot in different hardware infrastructure for any of the pipeline stage.

E.g. Baking can be implemented by using dockers and scaled via kubernetes.

DEVOPS SIDE

I will develop a powershell script that will take a json document and create / update databricks environment from scratch and other components, based on the configuration provided.

```
{
    “eventhubs”: [
        {
            “material”: “dough”,
            “expiresAfter”: “3hrs”,
            “storageTemprature” : “22C”
        }
    ]

    “databricks”: {
        “environment”: “production”,
        “jobs” : [
            {
                “notebook”: “CookieOven/mixing”,
                “cluster”: “InductionOptimizedClusters”,
                “workers”: 8
            }
        ]
    }
}
```