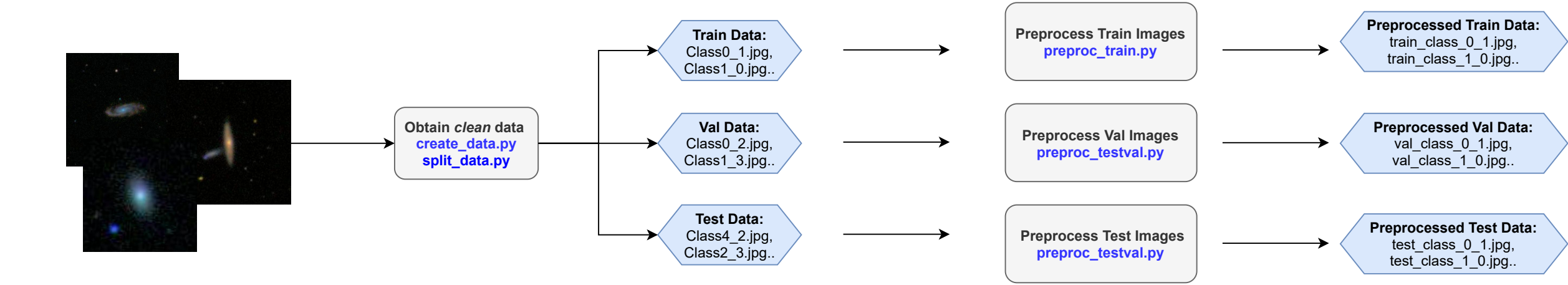


Data Acquisition	Clean Data	Data Preprocessing
Data is downloaded and unzipped in cwd	Reduce to 28790 images. 8434, 8069,578, 3903, 7806	Separate jobs for train data, and test/val data
Pegasus		

Download **galaxy-zoo-the-galaxy-challenge.zip**
File size: 1.9 GB

The images are assigned to classes based on their properties. Subset of the images is used as a dataset going forward. Total 27,675 images each between 8 kB to 16 kB. **Total: 386.0 MB**



Pegasus

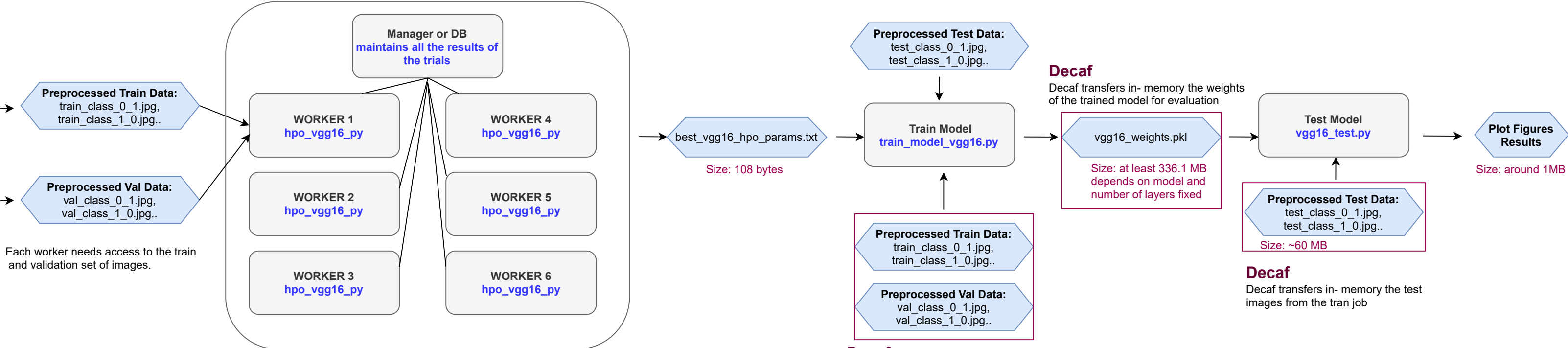
Downloads the data from an online repository.
1.9 GB

The data preprocessing jobs can be parallized based on number of available nodes. **Total Input: 386.0 MB**

Pegasus

Parallelizes the preprocessing jobs.

Hyperparameter Optimization	Train Model	Test Model
Tune model to find the best setting (Optimizer, Activation)	Use the best optimization parameters to train model	7 classification metrics accuracy, precision, recall, F1, confusion matrix, ROC, AUC
Decaf		



Each worker needs access to the train and validation set of images.

Pegasus

Lunches one big HPO job inside which Decaf runs communication between nodes

Decaf

Decaf transfers in- memory all the train and validation images from one of the HPO jobs

Decaf

Decaf transfers in- memory the weights of the trained model for evaluation

Decaf

Decaf transfers in- memory the test images from the tran job