Project Summary
Head Start Data Lake Feasibility Evaluation and Pilot Test
with support from the Bill & Melinda Gates Foundation

Goal

The National Head Start Association (NHSA) seeks to create an early childhood data lake, a central repository that will collect, store, and analyze data from Head Start programs in ways that help Head Start programs learn from each other. Our data lake will be designed with a primary goal of providing useful information back to programs and partners in order to improve early childhood practice and outcomes.

Our intention is to bring together data from Head Start programs nationwide to provide actionable information about children, families, staff, and programs. The data lake will improve systems and drive practical improvement. This data will be organized and reported back to programs to inform their daily practice and help them systematically improve their programs and services. The aggregated data will enable NHSA to identify national trends needed to support advocacy efforts, inform policy recommendations, and drive professional development for the Head Start field.

Our Model

Currently, while Head Start programs generate a significant amount of data about children, families, staff, communities, and operations, it is difficult for many to use this data to inform planning and improve practice. The barriers come in many forms, from technological limitations to staff capacity to backwards motivational mechanisms. Data has also historically been used to punish programs; it has been a hammer, not a flashlight. We seek to change that.

NHSA will take a user-centered design approach to this work, putting the experiences and needs of the Head Start practitioner at the center of the work. At its simplest, this project rests on the belief that user-driven solutions result in more active engagement, higher quality engagement, and ultimately better outcomes than externally-developed and/or regulatorily-mandated solutions.

We firmly believe, by driving this effort from the practitioner perspective, programs will feel safe enough to participate (meaning put their data in the lake) and engage in improvement efforts without fear. A national effort managed by NHSA will ensure this is available to all programs, no matter their size, budget, or technical capacity. This model will also allow programs to ask questions, propose additional data points and analyses, and help us continuously expand and improve the data lake to support program practice and improvement.

Project Plan
Phase 1

Before we can begin the technical work, we need to accomplish three things: (1) set up the legal structure governing the data lake; (2) find technical experts; and (3) identify the initial data to be included. We have set up each of these tasks intentionally to support the long-term sustainability of the data lake:

1. **Legal structure**. Head Start programs must feel safe engaging with the data lake, meaning they are sure their data is safe and secure and none of the information will be used to harm their children, families, staff, or program's existence as a whole. We will host two meetings with Head Start practitioners (4-6 pilot programs), expert stakeholders, and legal counsel during which we will develop a data sharing agreement aimed at all the protections needed from a program perspective as our priority. Because we have plans to expand the pilot, we will also establish a governance structure (such as a review board) that will be responsible for making decisions to ensure the data lake remains true to its intended purpose.

2. **Technical experts**. Through a formal process, we will select a technical contractor who will create and maintain the infrastructure of the data lake. The data lake will require significant resources to maintain long-term, and it is important that it is accessible without unfairly draining resources from participating programs. To ensure long term sustainable and equitable access, the contractor will help us conduct a feasibility study and develop a business plan.

3. **Initial data**. For the pilot, we will start with a set of data points NHSA has been working on that we believe can be:
   a. collected by all Head Start programs;
   b. reviewed regularly to help leaders understand the health of their organization; and
   c. could serve as warning signs for potential problems before they happen.
   We are starting with this set because we believe they will provide significant benefit to all programs who participate, and valuable and easily-usable data is key to scaling and sustainability. While we will start the data lake with this discrete set of data points we can share back with programs in a useful way, we anticipate programs will ask questions, want to slice and dice the data, and want additional data points once they start using the data lake on a regular basis.

   The data lake could expand to include data from other systems, such as state and local pre-K and K-12 school systems, child care providers, and more, but the intended use of the data lake (providing useful information back to programs and partners in order to improve early childhood practice and outcomes) is our primary goal. This data will not be made available to federal, state, or local government systems for the purpose of monitoring, licensing, evaluation, or similar.

Phase 2

Building a successful data lake requires two pieces of work in addition to the groundwork laid out above: (1) the highly technical component of actually standing up the infrastructure and connecting to external data sources; and (2) designing the user experience to be useful and effective.

1. **Technical Components**. With the help of the technical contractor, NHSA will stand up the data lake infrastructure. Then, we will establish a pipeline to the ChildPlus Software System (used by the majority of the Head Start field) so data can flow directly from ChildPlus to the data lake. We will begin by pulling in the set of data points identified in Phase 1, but building this pipeline sets us up well for expansion, because it will then be much easier for other ChildPlus users to join in the future and for us to pull in additional data points as needed. After the pilot, we can also establish pipelines to the other database systems used by Head Start programs using the same model.

2. **User Experience**. We will focus on ensuring the data we bring into the data lake is organized and presented back to programs in a useful way. We will work closely with the pilot participants to build and iterate on dashboards and usable reports to inform their daily practice. We will also work on an aggregated dashboard for national analysis and peer-to-peer learning with adequate safeguards in place for programs.

This document is a topline summary of the Head Start Data Lake Feasibility Evaluation and Pilot Test. We hope this provides a clear picture of our intentions and the general steps we will take to achieve them. Questions? Email Victoria Jones: vjones@nhsa.org